**Journal of Law and Policy**

Volume 14
Issue 2
SCIENCE FOR JUDGES VI:
Techniques for Evidence-Based Medicine

Article 4

2006

# Evaluating Systematic Reviews and Meta-Analyses

Lisa A. Bero

# EVALUATING SYSTEMATIC REVIEWS AND META-ANALYSES

*Lisa A. Bero, Ph.D.**

> If one's first impression of the world's clinical literature is that of its fearsome immensity; one's second is likely to be that of its appallingly poor average quality. The two are obviously interconnected; the drug literature is overburdened by a vast volume of superfluous and even dangerous rubbish. The standards of medical journals range from the sublime (of which there are very few) to the disgraceful.
>
> Graham Dukes, MD MA LLM, 1977[1]

## INTRODUCTION

Information overload in the medical field is not a new problem. In fact, it just gets worse as more information of questionable validity accumulates. As much of this information appears as "scientific evidence" in the courtroom, there is a pressing need for law professionals to understand state-of-the-art methods for

* Professor in the Department of Clinical Pharmacy at the Institute for Health Policy Studies, University of California, San Francisco. Dr. Bero thanks Nick Royle, CEO, Cochrane Collaboration for assistance with preparing figure 2.

[1] Dr. Graham Dukes is currently with the Unit for Drug Policy Studies, University of Oslo. Having held senior positions with national regulatory authorities, World Health Organization (WHO), and World Bank (WB), he has a distinguished professional record. Graham Dukes has peerless experience in the areas of drug policy, legislation, regulation, utilization studies, information services, adverse reaction monitoring, medical risk management and professional training. He has assisted numerous countries in the development of new polices, reorganization of regulatory systems, the design of new pharmaceutical legislation and supply structures.

critically appraising and summarizing massive amounts of scientific data.

Systematic reviews and meta-analyses are powerful methods for gathering, critiquing and summarizing medical and scientific information. Systematic reviews are combinations of results that adhere to pre-defined methods, but that may not result in quantitative combination of the data. Meta-analysis is a quantitative approach to systematically combining the results of previous studies. A meta-analysis that does not start as a systematic review may be published, but, for reasons described below, it would not be a high quality review.

The use of meta-analyses and systematic reviews to guide clinical practice is increasing.[2] Systematic reviews and meta-analyses of randomized controlled trials are the most methodologically rigorous forms of evidence to evaluate the effectiveness of therapeutic interventions, particularly pharmacotherapy.[3] They often form the foundation for practice guidelines, clinical decision support systems, drug formulary decisions, and drug payment schemes.

Healthcare practitioners, researchers, and anyone interested in answering scientific or medical questions would like to be able to turn to the ideal report of research findings. This ideal report would, in one place, summarize data from all studies available on a particular topic. The studies would be critically evaluated using unbiased methods. The report would be instantly accessible and kept up to date as new data accumulated.

The Cochrane Collaboration, whose logo is illustrated in Figure 1,[4] aspires to provide this ideal source of information. The Cochrane Collaboration was founded in 1993 and named after

---

[2] Lisa A. Bero & Drummond Rennie, *The Cochrane Collaboration: Preparing, Maintaining, and Disseminating Systematic Reviews of the Effects of Healthcare*, 274 JAMA 1935, 1935-38 (1995); Lisa A. Bero & Alejandro R. Jadad, *How Consumers and Policymakers can use Systematic Reviews for Decision Making*, 127 ANN. INTERNAL MED. 37, 37-42 (1997); Deborah J. Cook et al., *Systematic Reviews: Synthesis of Best Evidence for Clinical Decisions*, 126 ANN. INTERNAL MED. 376, 376-80 (1997).

[3] Cook et al., *supra* note 2.

[4] *See infra* note 65.

British epidemiologist, Archie Cochrane. In 1979, Cochrane wrote that "[i]t is surely a great criticism of our profession that we have not organized a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomized controlled trials."[5] The long term goal of the Cochrane Collaboration is to develop these critical summaries of all trials of all healthcare interventions. The Cochrane Collaboration is an international non-profit organization that aims to help people make well-informed decisions about healthcare by preparing, maintaining and promoting the accessibility of systematic reviews of the effects of healthcare interventions. It produces and disseminates systematic reviews of healthcare interventions and promotes the search for evidence in the form of clinical trials and other studies of interventions.[6] The major product of the Collaboration is the *Cochrane Database of Systematic Reviews*, which is published quarterly as part of the Cochrane Library.[7]

---

[5] Archie L. Cochrane, *1931-1971: A Critical Review, with Particular Reference to the Medical Profession*, *in* LONDON: OFFICE OF HEALTH ECONOMICS, MEDICINES FOR THE YEAR 2000 1-11 (1979).

[6] The Cochrane Collaboration Home Page, http://www.cochrane.org (last visited May 9, 2006).

[7] *See infra* note 65. The Cochrane logo (Figure 1) is a meta-analysis of 7 randomized controlled trials comparing a short, inexpensive course of a corticosteroid to placebo in women with premature labor. The data for these trials are shown as "odds ratios," meaning the odds of patients in the treated group having the outcome divided by the odds of the patients in the placebo group having the outcome. An odds ratio is just one way of representing the point estimate, or result, of a trial. Each horizontal line represents the results of one trial; the middle of the line is the point estimate for the odds ratio and the ends of the line represent the variability around this estimate. The shorter the line, the larger the trial and the more certain the result. In the case of the Cochrane logo, an odds ratio of less than one will show a favorable effect of the treatment because this would mean that the odds of a woman in the corticosteroid group having a baby die from complications of premature labor would be less than the odds of a woman in the placebo group. As seen in the figure, the seven trials all yield slightly different point estimates or odds ratios.
The diamond represents the statistical combination of the results of all seven trials. The vertical line indicates an odds ratio of one, or the position at which there would be no difference in outcome between the treated and control groups. If a horizontal line touches the vertical line, it means that that particular trial

The validity of a systematic review depends on the extent to which the methods of the review reduce random error and systematic bias. Systematic reviews reduce bias because they are conducted according to strictly defined methods. A good systematic review contains a focused question, an explicit and comprehensive search strategy, explicit inclusion and exclusion criteria that are uniformly applied, a rigorous critical appraisal of each identified study and, if appropriate, a quantitative summary of the evidence.[8]

Part I of this paper discusses the importance of systematic reviews and meta-analyses for evidence-based medicine by highlighting the benefits systematic reviews provide researchers and funders over reliance on individual studies. Part II offers a primer on how to evaluate a systematic review to best eliminate bias in the review and provides some general guidelines for each step in the optimal review process.

## I. THE IMPORTANCE OF SYSTEMATIC REVIEWS

Systematic reviews are a way of dealing with the massive information overload that is typical of clinical medicine. Systematic reviews are an efficient scientific technique for gathering, critiquing and summarizing large amounts of information. A systematic review allows the reader to see when scientific findings are consistent. When studies that are done in slightly different ways or in slightly different populations reach the same answer, we can assume that the results may be generalizable to a wider population. On the other hand, systematic reviews allow for the exploration of inconsistencies and conflicts in the results of individual studies. By presenting the same information on all studies in the systematic reviews, the review allows the reader to determine whether divergent results might be due to differences in

found no clear difference between the treatments. The position of the diamond to the left of the vertical line indicates that the treatment studied is beneficial. The logo demonstrates that corticosteroid therapy for women in premature labor is an effective intervention.

[8] Cynthia D. Mulrow, *The Medical Review Article: State of the Science*, 106 ANN. INTERNAL MED. 485 (1987).

the methods of the original studies, differences in the experimental intervention tested, or variability in the characteristics of the populations tested.

Increasing the power, or sample size, of a study reduces random error. The larger the study, the more likely that the results will be distributed around the true effect. In smaller studies, due to chance and random error, the results are less likely to represent the true effect. Meta-analysis increases the power of a study because it combines the results of a number of small studies into one study with a larger sample size. Thus, meta-analysis increases the precision of an estimate of an effect by decreasing the variability around the estimate as the sample size increases.

A common myth about meta-analysis is that if enough studies are combined, the results will always be statistically significant, or demonstrate an effect. This is not the case, however, as illustrated by a meta-analysis of randomized, controlled trials of prophylactic lidocaine for acute myocardial infarction.[9] In this meta-analysis, the results of eight small randomized controlled trials of lidocaine were combined. When the studies were statistically combined into a meta-analysis of almost 9,000 patients, the summary estimate remained statistically non-significant, thus showing there was no effect of lidocaine.

Another valuable contribution of systematic reviews is that they can help set research agendas (and avoid embarrassment of researchers) by identifying what questions have been answered, as well as gaps in understanding. This contribution is best illustrated with the technique of cumulative meta-analysis, which means that each study in the meta-analysis is added in consecutively.[10] Figure 2,[11] represents the results of a cumulative meta-analysis of randomized controlled trials examining the effect on mortality of

---

[9] Elliott M. Antman et al., *A Comparison of Results of Meta-analyses of Randomized Control Trials and Recommendations of Clinical Experts: Treatments for Myocardial Infarction*, 268 JAMA 240, 242-44 (1992).

[10] Joeseph Lau et al., *Cumulative Meta-analysis of Clinical Trials Builds Evidence for Exemplary Medical Care*, 48 J. CLIN. EPIDEMIOL. 45, 45-47, 59-60 (1995).

[11] *See infra* note 66.

thrombolytic therapy following an acute myocardial infarction.[12] As shown in the figure, after the enrollment of approximately 6,000 patients in 27 trials over a 20 year period, it was clear that treatment with thrombolytic therapy reduced mortality. As more patients were enrolled in more trials, the odds ratio did not change much, although the variability around the odds ratio decreased as the sample size of the cumulative meta-analysis increased. Thus, approximately 42,000 people participated in 43 more trials that were not needed to determine if a thrombolytic therapy effectively reduces mortality after a myocardial infarction. Meta-analysis is a valuable tool for helping researchers and funders decide when more research is needed to answer a clinical question.

If a good systematic review is available, should one rely on the results of an individual study to answer a clinical question? The answer is no. Systematic reviews and meta-analyses have become the gold standard for evidence-based medicine. Evidence-based medicine emphasizes the examination of evidence from clinical research over intuition, unsystematic clinical observations, and pathophysiological rationale for clinical decision making.[13] Because systematic reviews and meta-analyses of randomized controlled trials are designed to reduce bias and generalize results across patients, they are considered to be the top of the "hierarchy of evidence" for evidence-based medicine.[14] Conclusions drawn from a single randomized controlled trial are generally considered weaker because they are based on smaller sample size and do not generalize across different patients.[15] The remainder of this paper will present some general guidelines for evaluating the validity of

---

[12] Antman, *supra* note 9, at 240-48. Joseph Lau et al., *Cumulative Meta-analysis of Therapeutic Trials for Myocardial Infarction*, 327 NEW ENG. J. MED. 248, 251 (1992).

[13] David L. Sackett & W.M. Rosenberg, *The Need for Evidence-based Medicine*, 88 J. ROYAL SOC'Y MED. 620, 620-24 (1995); David L. Sackett, *Evidence-based Medicine*, 21 SEMINARS IN PERINATOLOGY 3, 3-5 (1997).

[14] Gordon H. Guyatt et al., *Users' Guides to the Medical Literature: IX. A Method for Grading Healthcare Recommendations*, 274 JAMA 1800 (1995); Cook et al., *supra* note 2.

[15] Joseph C. Cappelleri et al., *Large Trials vs. Meta-analysis of Smaller Trials: How Do Their Results Compare?*, 276 JAMA 1332, 1332-38 (1996).

systematic reviews.

## II. EVALUATING SYSTEMATIC REVIEWS

There are many possible sources of bias in reviews, including the framing of the research question, the selection of studies to be included, the extraction of data from and critical appraisal of the included studies, and the analysis. It is sometimes difficult to detect the source of bias. For example, reviews of studies of adverse effects related to exposure to secondhand smoke are more likely to conclude that secondhand smoke is not harmful when the authors of the reviews are affiliated with the tobacco industry, regardless of the methodological quality of the review, whether it was published in a peer-reviewed journal, or other factors.[16]

One way to avoid bias in a review is to develop a protocol for the review before commencement and adhere to the protocol regardless of the results of the review.[17] A reader of the review can then determine whether the authors conducted it according to the systematic methods proposed. When the reviews are completed, the readers can be assured that the authors adhered to the methods of the protocol and did not change the methods after they started the review. Adherence to a strict protocol can sometimes result in reviews where no studies that meet the criteria for the review can be found. However, as mentioned above, these reviews are still useful for identifying gaps in the research literature.

A good protocol (and completed review) should contain the following sections: 1) an objective or research question, 2) criteria for selecting studies for the review, 3) a search strategy for studies, 4) methods for assessing the validity of included studies, 5) methods for selecting studies for the review, 6) methods for collecting data from the studies, and 7) an analysis plan.

---

[16] Deborah E. Barnes & Lisa A. Bero, *Why Review Articles on the Health Effects of Passive Smoking Reach Different Conclusions*, 279 JAMA 1566, 1566-70 (1998).

[17] The Cochrane Library publishes protocols of reviews.

*1. Objectives*

Every systematic review should contain a precise statement of the primary objective or research question. The objective should include a description of the population to be tested, the intervention or exposure to be tested, the treatment for the comparison group, and the outcome.[18] In other words, "What is being tested to change what outcome in whom?"[19]

Any prior hypotheses and comparison groups should be stated. This includes pre-specified subgroup analysis. For example, one might hypothesize that a drug will reduce hypertension only in non-obese patients. Thus, the review should contain an objective stating that data from obese patients will be analyzed separately from data from non-obese patients. Specifying subgroup analysis after data collection for the review has already begun can be a "fishing expedition" or "data dredging" for statistically significant results and is not appropriate.

*2. Criteria for Selection of Studies for the Review*

The biased citation of studies in a review can be a major source of error in the results of the review.[20] Authors of reviews can influence their conclusions by citing only studies that support their preconceived, desired outcome. The clearly stated objective of the review determines the criteria used to select studies for inclusion in

---

[18] Andrew D. Oxman et al., *Users' Guides to the Medical Literature: VI. How to Use an Overview*, 272 JAMA 1367, 1367-71 (1994).

[19] Some examples of clear objectives are:

- Do corticosteroids, compared to no treatment, prevent pre-term labor in pregnant women?
- Does prophylactic lidocaine, compared to placebo, prevent acute myocardial infarction in patients who have already had a myocardial infarction?
- Do calcium channel blockers lower blood pressure in patients with hypertension compared to placebo? A related question with a different outcome of interest would be, "Do calcium channel blockers reduce mortality in patients with hypertension compared to placebo?"

[20] Paul F. Neihouse & Susan C. Priske, *Quotation Accuracy in Review Articles*, 23 DICP 594, 594-96 (1989).

the review. The eligibility criteria should include a description of 1) the participants in the studies (e.g., children, men or women with recent heart attacks), 2) the interventions or exposures (e.g., a drug, chemical exposure), 3) the outcome measures (e.g., mortality, heart attack), and 4) the study design (e.g., randomized controlled trial, observational study). As mentioned above, systematic reviews of randomized controlled trials are considered one of the most rigorous types of clinical evidence. In a randomized controlled trial, the only difference between the two groups being compared is the experimental intervention. Thus, systematic reviews, particularly those examining the effects of therapeutic interventions, may include only randomized controlled trials.

The types of study designs to be included in a review will vary with the research objective. For example, the effects of environmental toxins are not typically examined using randomized controlled trials. Thus, a systematic review of the effects of a potential environmental hazard will include studies of observational designs, such as cohort or case control studies.[21] Even qualitative studies, such as focus groups or interview studies, can be combined using systematic review methods. For example, a review of qualitative studies on barriers to childhood vaccination identified several consistent areas that are obstacles to children receiving immunizations.[22]

*3. Search Strategy for Identification of Studies*

After the inclusion criteria for studies are clearly specified, a comprehensive search strategy must be developed.[23] The search should be as comprehensive as possible. The review should specify

---

[21] *See* Barnes & Bero, *supra* note 16.

[22] Edward Mills et al., *Systematic Review of Qualitative Studies Exploring Parental Beliefs and Attitudes Towards Childhood Vaccination Identifies Common Barriers to Vaccination*, 58 J. CLIN. EPIDEMIOL. 1081, 1081-88 (2005).

[23] Carl Counsell & Hazel Fraser, *Identifying Relevant Studies for Systematic Reviews*, 310 BMJ 126 (1995); Maureen O. Meade & W. Scott Richardson, *Selecting and Appraising Studies for a Systematic Review*, 127 ANN. INTERNAL MED. 531, 531-37 (1997).

the exact dates of the search and whether there were any language restrictions.

Unfortunately, a print review article is out of date as soon as it is published. Thus, regular updating of reviews, such as those in the Cochrane Library, is essential for ensuring the accuracy of the information. Although the end date of a search for studies should be as recent as possible (and regularly updated), the start date of the search should be appropriate to the question. For example, if no trials of a particular drug were conducted before 1985, it is not necessary to extend the search prior to that date. In some cases, the circumstances under which research is conducted may change. For example, the definition of the AIDS diagnosis was refined during the late 1980's. Trials of HIV/AIDS therapies conducted in the early 1980's may have included different populations than those conducted in the 1990's. Thus, reviews of these therapies should clearly specify the rationale for the search dates.

Many systematic reviewers restrict their searches to English language-only studies. However, this is primarily for the sake of convenience and can introduce a number of limitations. The methodological quality of clinical trials does not vary by language of the publication, so quality concerns are not a good justification for language restrictions.[24] Furthermore, the results of systematic reviews can change completely when the review includes only English language studies or studies in any language. Gregoire found that in at least one out of 36 consecutive meta-analyses the exclusion of papers for language reasons produced results different from those which would have been obtained if this exclusion criteria had not been used.[25] As long as a study meets the inclusion criteria for the review, it should be included regardless of the language of its publication.

---

[24] D. Moher et al., *Completeness of Reporting of Trials Published in Languages Other Than English: Implications for Conduct and Reporting of Systematic Reviews*, 347 LANCET 363, 363-36 (1996); Matthias Egger et al., *Language Bias in Randomised Controlled Trials Published in English and German*, 350 LANCET 326, 326-29 (1997).

[25] G. Gregoire et al., *Selecting the Language of the Publications Included in a Meta-Analysis: Is There a Tower of Babel Bias?*, 48 J. CLIN EPIDEMIOL. 159, 159-63 (1995).

Electronic databases of research articles are a good starting place to search for studies that meet the inclusion criteria for a systematic review. MEDLINE or PubMed (produced for free by the National Library of Medicine) is the most commonly used database. However, a PubMed search identifies only about 50% of randomized controlled trials published in journals that are indexed by MEDLINE. Comparison of a MEDLINE search with a "gold standard" search based on manual, page-by-page searching of journals for randomized controlled trials found that MEDLINE was not good at detecting trials.[26] The poor performance of MEDLINE is due to the improper indexing of randomized controlled trials, as well as the inappropriate use of search terms.[27] Clearly, searching MEDLINE alone is inadequate for identifying randomized controlled trials. Supplementing the MEDLINE search with other electronic database searches can be useful. EMBASE indexes more than 100 journals that are not indexed in MEDLINE, including many non-English language journals, and LILACS is the largest electronic database of Spanish-language medical journals. Specialty electronic databases can also be useful, depending on the topic of the research questions. CINAHL, PsychLit, CancerLit and BIOSIS are examples of specialty databases.

As all electronic databases have limitations, a good review should employ additional methods for identifying studies that meet the inclusion criteria. Checking the reference lists of studies that are identified in the electronic searches often identifies additional studies. In addition, citation databases, such as Web of Science, can locate additional studies. These databases identify articles that cite the studies that were identified in the initial electronic search. Lastly, hand searching or the manual, page-by-page searching of journals for randomized, controlled trials is the gold standard for identifying studies. The Cochrane Collaboration has led an effort in hand searching journals to identify trials and these trials are indexed in the Cochrane Library.[28]

---

[26] Kay Dickersin et al., *Identifying Relevant Studies for Systematic Reviews*, 309 BMJ 1286, 1286-91 (1994).

[27] *Id.*

[28] *Id.*

Systematic reviewing and meta-analysis proceeds under the assumption that a complete and representative sample of relevant studies is available for analysis.[29] However, because access to relevant studies is frequently limited to *published* studies, systematic reviews and meta-analyses are particularly vulnerable to biases that may affect the publication of studies. The majority of methodologists and journal editors now believe that unpublished data should be included in systematic reviews, suggesting widespread belief that important data remain unpublished.[30]

The problem of publication bias, the tendency for studies showing statistically significant results to be published and published more quickly than studies with statistically non-significant results, poses a serious challenge to identifying studies for systematic reviews. First identified in 1959,[31] publication bias raises the concern that statistically significant study results may dominate the research record, thus reducing the range of evidence on which systematic reviews and meta-analyses are based.[32] A recent study modeling the probability of finding statistically significant findings that are not correct has concluded that "most published research findings are false."[33] Most studies also show

---

[29] Jerome M. Stern & R. John Simes, *Publication Bias: Evidence of Delayed Publication in a Cohort Study of Clinical Research Projects*, 315 BMJ 640, 640-45 (1997).

[30] Debra J. Cook et al., *Should Unpublished Data Be Included in Meta-analyses? Current Convictions and Controversies*, 269 JAMA 2749, 2749-53 (1993).

[31] Theodore D. Sterling, *Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance - or Vice Versa*, 54 J AM. STAT. ASS'N 30, 30-34 (1959).

[32] Kay Dickersin et al., *Publication Bias and Clinical Trials*, 8 CONTROL CLININCAL TRIALS 343, 343-53 (1987); Phillipa J. Easterbrook et al., *Publication Bias in Clinical Research*, 337 LANCET 867, 867-72 (1991); Anastasia L. Misakian & Lisa A. Bero, *Publication Bias and Research on Passive Smoking: Comparison of Published and Unpublished Studies*, 280 JAMA 250, 250-53 (1998).

[33] John P.A. Ioannidis, *Why Most Published Research Findings are False*, 2 PLOS MEDICINE 101, 101-06 (2005), *available at* http://medicine.plosjournals.org/archive/1549-1676/2/8/pdf/10.1371_journal.pmed.0020124-S.pdf.

that not only statistical significance but a large sample size is also associated with publication, so that small, statistically non-significant studies are rarely published. Thus, the results of systematic reviews and meta-analyses can be skewed in favor of new treatments showing positive initial results.[34] Publication bias poses a particular threat to the reliability and validity of systematic reviews and meta-analysis by leading to spuriously large treatment effects in early meta-analyses of the available evidence.[35]

Thus, authors of systematic reviews must attempt to identify unpublished and ongoing studies through a variety of methods. Searching the abstracts of conference proceedings, of which only about 50% are published as full journal articles,[36] is one mechanism for identifying unpublished studies. Personal communication with investigators who are active in the field is another method. Searching clinical trial registries is one of the most promising methods for identifying unpublished data.

Registration of clinical trials is one method that has been proposed to reduce publication bias.[37] The exposure of notable cases of data suppression from clinical trials prompted the International Committee of Medical Journal Editors and its 11 member journals to require, as a condition of consideration for publication, registration of clinical trials in a public trials registry.[38] Although extensive debate about the specific content of trial registries continues, the increasing availability of such registers will make it easier to identify unpublished studies for systematic reviews.

---

[34] John P.A. Ioannidis et al., *Issues in Comparisons Between Meta-analyses and Large Trials*, 279 JAMA 1089, 1089-93 (1998).

[35] *Id.*

[36] Roberta W. Scherer et al., *Full Publication of Results Initially Presented in Abstracts: A Meta-Analysis*, 272 JAMA 158, 158-62 (1994).

[37] Robert John Simes, *Publication Bias: The Case for an International Registry of Clinical Trials*, 4 J. CLINICAL ONCOLOGY 1529, 1529-41 (1986).

[38] Kay Dickersin & Yuan I. Min, *Publication Bias: The Problem that Won't Go Away*, 703 ANN. N.Y. ACAD. SCI. 135, 146-48 (1993); Catherine D. DeAngelis et al., *Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors*, 292 JAMA 1363, 1363-64 (2004).

*4. Assessment of Methods Used to Reduce Bias in the Selected Studies*

Bias is the combination of various design, data, analysis and presentation factors that tend to produce statistically significant research results that are not true.[39] Various factors can lead to statistically significant outcomes in randomized controlled trials of drug efficacy, including framing of the research question, design and analysis of the study, and conduct of the study.[40] Reporting (or not) of the full results or selective reporting of outcomes can also contribute to the problem of publication bias.[41] One criticism of meta-analyses and systematic reviews is: "Garbage in, garbage out." This means that if poorly designed and executed studies that fail to minimize bias are included in a systematic review, the results of the review will not be valid. Therefore, it is essential that the studies that are included in a systematic review are evaluated for their methodological quality—the methods used to reduce bias. The tools used to evaluate methodological quality should be specific to the study design being evaluated. As most evaluation tools have been developed to assess the quality of randomized controlled trials, the following section will focus on the evaluation of trials. However, instruments for assessing the methods of observational and qualitative studies are also available.[42]

*5. How Do We Measure Quality?*

Dozens of instruments for assessing the methodological quality

---

[39] Ioannidis, *supra* note 33.

[40] Cochrane Collaboration, *supra* note 6.

[41] *Id.*; Richard Smith, *Medical Journals are an Extension of the Marketing Arm of Pharmaceutical Companies*, 2 PLoS Med. 364, 364-66 (2005), *available at* http://medicine.plosjournals.org/archive/1549-1676/2/5/pdf/10.1371_journal.pmed.0020138-L.pdf.

[42] Andrew D. Oxman & David L. Sackett, *Users' Guides to the Medical Literature: I. How to Get Started. The Evidence-Based Medicine Working Group*, 270 J. AM. MED. ASS'N 2093, 2093-95 (1993); Mildred K. Cho & Lisa A. Bero, *Instruments for Assessing the Quality of Drug Studies Published in the Medical Literature*, 272 J. AM. MED. ASS'N 101, 101-04 (1994).

of randomized controlled trials exist,[43] including some frequently used instruments that contain from 3 to 22 items.[44] In 1995, Moher examined 25 published scales and 9 checklists for measuring the methodological quality of randomized controlled trials.[45] Most of these instruments calculate a quality "score" for the randomized controlled trial.

There are several problems with all of these quality assessment instruments. First, reliability and validity have not been measured for most of them. A reliability measurement would provide information on how often multiple coders, using the instruments independently, would derive the same score. A validity measure would provide information on whether the items assessed in the instrument are truly evaluating methods that reduce bias.

Second, most methodological quality assessment instruments combine the evaluation of reporting and actual study design. If two studies that are designed in an identical way are published in different journals, one may be reported more completely than the other. The more completely reported study would have a better quality score, although it is not truly a better designed study. In order to reduce the problem of variability in reporting, systematic reviewers often correspond with the authors of the studies to obtain information that is not in the study report. In addition, in recent years, many journals have developed reporting standards and have strengthened their policies regarding reporting of randomized controlled trials.[46]

A third problem with the quality assessment instruments is that there is little empirical evidence to support differential weighting of the individual components of the quality scores. Individual characteristics of randomized controlled trials that are associated

---

[43] *See* D. Moher et al., *Assessing the Quality of Randomized Controlled Trials: An Annotated Bibliography of Scales and Checklists*, 16 CONTROLLED CLINICAL TRIALS 62, 62-73 (1995); *See also* Moher, *supra* note 24.

[44] Thomas C. Chalmers et al., *A Method for Assessing the Quality of a Randomized Control Trial*, 2 CONTROLLED CLIN. TRIALS 31, 31-49 (1981); Cho & Bero, *supra* note 42.

[45] Moher, *supra* note 43.

[46] Colin Begg et al., *Improving the Quality of Reporting of Randomized Controlled Trials: The CONSORT Statement*, 276 JAMA 637, 637-39 (1996).

with bias have been identified in studies evaluating a variety of randomized controlled trials. These characteristics include inadequate randomization[47] and concealment of allocation.[48] Concealment of allocation means that investigators, at the beginning of the study and before any patients are assigned to treatment, are unaware of the group to which a patient will be randomly assigned. Other characteristics of randomized controlled trials that are associated with bias are inadequate double blinding,[49] insufficient sample size,[50] inappropriate choice of drugs to be compared,[51] and inappropriate choice of statistical analysis.[52] For

---

[47] Thomas C. Chalmers et al., *Controlled Studies in Clinical Cancer Research*, 287 NEW ENG. J. MED. 75, 75-78 (1972); Graham A. Colditz et al., *How Study Design Affects Outcomes in Comparisons of Therapy*, 8 STAT. MED. 441 (1989).

[48] Kenneth F. Schulz, *Subverting Randomization in Controlled Trials*, 274 JAMA 1456, 1456-58 (1995) [hereinafter Schulz, *Subverting Randomization*]; Kenneth F. Schulz et al., *Empirical evidence of bias. Dimensions of Methodological Quality Associated With Estimates of Treatment Effects in Controlled Trials*, 273 JAMA 408, 408-12 (1995) [hereinafter Schulz, *Empirical Evidence*].

[49] Colditz, *supra* note 47; Schulz, *Empirical Evidence*, *supra* note 48.

[50] John P. Ioannidis, *Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials*, 279 JAMA 281, 281-86 (1998); Misakian & Bero, *supra* note 32; Bodil Als-Nielsen et al., *Association of Funding and Conclusions in Randomized Drug Trials: A Reflection of Treatment Effect or Adverse Events?*, 290 JAMA 921, 921-28 (2003); John P. Ioannidis et al., *Randomised Trials Comparing Chemotherapy Regimens for Advanced Non-small Cell Lung Cancer: Biases and Evolution Over Time*, 39 EUR. J. CANCER 2278, 2778-87 (2003); Bodil Als-Nielsen et al., *Are Trial Size and Quality Associated with Treatment Effects in Randomised Trials*?, 12TH ANNUAL COCHRANE COLLOQUIUM (2004).

[51] P.A. Rochon, et al., *A Study of Manufacturer-Supported Trials of Nonsteroidal Anti-inflammatory Drugs in the Treatment of Arthritis*, 154 ARCHIVES INTERNAL MED. 157, 157-63 (1994); Helle Krogh Johansen & Peter C. Gotzsche, *Problems in the Design and Reporting of Trials of Antifungal Agents Encountered During Meta-analysis*, 282 JAMA 1752, 1752-59 (1999); Benjamin Djulbegovic et al., *The Uncertainty Principle and Industry-Sponsored Research*, 356 THE LANCET 635, 635-38 (2000); Daniel Safer, *Design and Reporting Modifications in Industry-Sponsored Comparative Psychopharmacology Trials*, 190 J. OF NERVOUS AND MENTAL DISEASE 583, 583-92 (2002).

example, Schulz and colleagues found that estimates of treatment effects were exaggerated by 41% for inadequately concealed trials and by 17% for trials with inadequate double blinding.[53]

Most quality assessment instruments assign points for using appropriate methods for each component of the randomized controlled trial, and sum these points into the quality score. Recent research suggests that these scores are not valid measures of methodological quality. Juni and colleagues determined that the use of different quality assessment scales using summary scores resulted in different conclusions of meta-analytic studies and proposed that specific components of methodological quality (e.g., concealment of allocation, blinding) should be individually assessed.[54] Therefore, the use of quality scores to rate the studies included in a meta-analysis should be viewed with caution. Reporting each included study's performance on the individual components of the quality score is more informative.

Some systematic reviewers also report additional characteristics of included studies that are not strictly measures of methodological quality. For example, peer-reviewed journal articles are less likely to have statistically significant outcomes than non-peer-reviewed journal articles.[55] In addition, a large body of evidence suggests that industry sponsorship of research is also associated with statistically significant results that are favorable to the sponsor. Two recent systematic reviews identified 19 studies examining the association of industry sponsorship and research outcomes.[56] The magnitude of this observed association is variable.

---

[52] Oscar H. Brook et al., *Effects of Coaching by Community Pharmacists on Psychological Symptoms of Antidepressant Users: A Randomised Controlled Trial*, 13 EUR. NEUROPSYCHOPHARMACOLOGY 347, 347-54 (2003); Jorge Gomez Cerezo et al., *Outcome trials of COX-2 Selective Inhibitors: Global Safety Evaluation Does Not Promise Benefits*, 59 EUR. J. CLINICAL PHARMACOLOGY 169, 169-75 (2003).

[53] Schulz, *Subverting Randomization*, *supra* note 48.

[54] Peter Juni et al., *The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis*, 282 JAMA 1054, 1054-60 (1999).

[55] Mildred K. Cho & Lisa A. Bero, *The Quality of Drug Studies Published in Symposium Proceedings*, 124 ANN. INTERN. MED. 485, 485-89 (1996).

[56] Justin E. Bekelman et al., *Scope and Impact of Financial Conflicts of*

For example, review articles on the health effects of secondhand smoke exposure that are written by tobacco industry-supported authors are about 90 times more likely to conclude that secondhand smoke is not harmful than those that are written by authors not affiliated with the tobacco industry.[57] Pharmaceutical industry-sponsored drug studies are about 4 times more likely to have conclusions that favor the sponsor than those that are funded by non-pharmaceutical sponsors.[58] Financial ties of investigators to their sponsors (e.g., stock ownership, consulting income, honoraria) are also associated with favorable research outcomes for the sponsor.[59]

In summary, a systematic reviewer should use common sense measures to assess the methods of studies that are included in the review. The components that are assessed for each included study should focus on the key features of the study, should be empirically verified to influence outcome, and should be reported individually. Systematic reviewers must keep in mind that the evaluation of included studies could indicate that all the studies are flawed. In this case, no conclusions should be drawn from the studies included in the review.

### *6. Methods for Selecting Studies for the Review, Extracting Data, and Appraisal of Studies*

Bias can be introduced during the selection of studies for inclusion in the review, as well as during the extraction and appraisal of data from the studies. Rigorous systematic reviewers often use two coders to independently select the studies from the list generated by the comprehensive search. The study selection should be done according to an explicit, written list of inclusion criteria. The coders should keep a written record of which criteria are met by each included study. The systematic review should also

---

*Interest in Biomedical Research: A Systematic Review*, 289 JAMA 454, 454-65 (2003); Joel Lexchin et al., *Pharmaceutical Industry Sponsorship and Research Outcome and Quality: Systematic Review*, 326 BMJ 1167, 1167-70 (2003).

[57] Barnes & Bero, *supra* note 16, at 1566-70.

[58] Lexchin, *supra* note 6.

[59] *Id.*

include a table of excluded studies listing the reasons why each study was excluded. Such a table is useful for determining if bias was introduced into the study selection process.

Two coders should also independently extract data from each included article and perform the quality assessment. Coders should be trained to use a data extraction form and be provided with a comprehensive set of instructions. Studies are sometimes assessed in random order using a computer random number generator in order to avoid the "training" effect that occurs as coders become more familiar with the data extraction instrument.

Systematic reviews sometimes report the inter-rater agreement among multiple coders. A higher degree of agreement gives the reader more confidence that the selection and data extraction process did not introduce bias into the review. Often disagreements between coders can be resolved by consensus. In these cases, the consensus is often reported in the review.

Reviewers are sometimes masked or blinded to the authors, authors' institutions, or journals in which studies are published. However, complete blinding is difficult to achieve. Furthermore, there is no evidence that the masking of the data coders influences their selection or quality ratings.[60] Therefore, it is less common for more recent systematic reviews to mask reviewers to the identity or sources of the studies.

In summary, a systematic review should include a description of how the data collection was done and whether it is complete. The review should report that pre-tested, standardized data collection forms were used by multiple coders working independently. The methods to resolve disagreements among the coders should be described. In addition, efforts to obtain information that was missing from the original study reports should be described. For each study included in the review, data should be reported on: methods (including study design, individual

---

[60] A.R. Jadad et al., *Assessing the Quality of Reports of Randomized Clinical Trials: Is Blinding Necessary?*, 17 CONTROLLED CLINICAL TRIALS 1, 1-12 (1996); J.A. Berlin, *Does Blinding of Readers Affect the results of Meta-Analyses? University of Pennsylvania Meta-analysis Blinding Study Group*, 350 LANCET 185, 185-186 (1997); M. Clarke et al., *Individual Patient Data Meta-Analysis in Cancer*, 77 BRITISH J. OF CANCER 2036, 2036-44 (1998).

quality assessment components), participants, interventions, outcomes, and results (recorded in natural units and converted to a common effect size when possible).

*7. Data Synthesis—A Meta-analysis or Not?*

The last step in conducting a systematic review is to decide whether the data resulting from the search, data extraction and critical appraisal should be summarized quantitatively into a meta-analysis. Data from the included studies should be quantitatively combined into a point estimate only if the participants, interventions, and outcomes are sufficiently similar. Although there are statistical techniques available to assess the heterogeneity of studies, deciding whether to combine results is largely a judgment call.[61] Data from individual studies should never be statistically combined if no studies of good methodological quality exist or if a very broad question is being addressed. Thus, it is acceptable to combine oranges and oranges or apples and apples. Although it is not acceptable to statistically combine apples and oranges, it is acceptable to do a systematic review of apples and oranges as long as one is interested in fruit. For example, if a reviewer is interested in the efficacy of continuing medical education (CME) to change physician behavior, she will gather studies that have tested a variety of educational methods, such as lecture, problem-based small groups, or online courses. The reviewer could include all of these studies in a systematic review in order to get an overview of how CME is conducted. However, she should only statistically combine studies that tested the same intervention, i.e., lectures or small groups.

"Vote counting" is not a valid method for summarizing the results of a systematic review. For example, if a review includes 11 studies, one might conclude that the intervention is effective if 7 studies found a statistically significant effect of the intervention and 4 did not. This vote count, however, negates many of the strengths of the systematic review technique, such as giving more

[61] JOSEPH L. FLEISS ET AL., STATISTICAL METHODS FOR RATES AND PROPORTIONS 161-165 (John Wiley & Sons Inc. 3d ed. 2003).

weight to studies of better methodological rigor. If the most poorly designed studies are those that show the significant effect, the conclusion that the intervention works may be erroneous. Thus, statistical combination of data from similar studies allows for the weighting of the studies according to their design characteristics, sample size, or other features that might affect outcome. If a meta-analysis is to be conducted, reviewers need to decide how the effect of the intervention examined in each study will be summarized. A discussion of the appropriateness of different summary statistics for a meta-analysis is beyond the scope of this paper.[62] However, the statistical methods used to combine data for meta-analyses do not differ in principal from those used in primary research. Parametric, non-parametric, regression and Bayesian techniques can be used.[63] The statistic chosen should be appropriate to the type of data analyzed and the reasons for choosing the statistic should be transparent. For example, dichotomous data, such as mortality, may be summarized as an odds ratio, relative risk, absolute risk difference, and number needed to treat.[64] Continuous data, such as blood glucose or blood pressure, can be combined directly if measured on the same scale, or converted to a common metric if measured on different scales. Although different statistical methods are used to combine data from observational studies, the principles of combining similar studies and exploring reasons for heterogeneity among studies are the same.[65]

---

[62] *See* Joseph Lau et al., *Quantitative Synthesis in Systematic Reviews*, 127 SYSTEMATIC REV. SERIES 91, 91-101, (Cynthia Mulrow, MD, MSc & Deborah Cook MD, MSc eds., 1997).

[63] J. L. Fleiss, *The Statistical Basis of Meta-Analysis*, 2 STAT. METHODS MED. RES. 121, 121-45 (1993); I. Olkin, *Statistical and Theoretical Considerations in Meta-Analysis*, 48 J. CLIN. EPIDEMIOL. 133, 147 (1995); T.C. Smith et al., *Bayesian Approaches To Random-Effects Meta-Analysis: A Comparative Study*, 14 STAT. MED. 2685, 2685-99 (1995).

[64] J.C. Sinclair & M. B. Bracken, *Clinically Useful Measures of Effect in Binary Analyses of Randomized Trials*, 47 J. CLIN. EPIDEMIOL. 881**,** 881-89 (1994).

[65] S. Greenland & M. P. Longnecker , *Methods for Trend Estimation from Summarized Dose-response Data, with Applications to Meta-analysis*, 135 AM J. EPIDEMIOL. 1301, 1301-09 (1992); W. Dumouchel, *Meta-analysis for Dose-*

As described in the section on identifying studies, a good systematic reviewer will attempt to control for publication bias by conducting a comprehensive search for ongoing and unpublished studies. However, most meta-analyses also contain a statistical estimate of publication bias.[66] These estimates tell the reader of the review whether publication bias exists among the studies included in the review and whether imputed results from unpublished studies might change the result of the review.

*H. Sensitivity Analyses*

Regardless of the statistical method used to combine data, sensitivity analyses should be conducted to measure the robustness of the summary effect. Variation in the characteristics of patients, interventions and study design features is inevitable across different types of studies. Therefore, it is important to explore whether any variation in the outcomes of the studies are due to these expected differences. A sensitivity analysis determines whether the summary point estimate is influenced by the assumptions made in conducting the systematic review. For example, a sensitivity analysis could be conducted in which the summary statistic is calculated first using all included studies, then recalculated after studies with certain characteristics are deleted from the analysis. If the results of the meta-analysis remain consistent, one has more confidence in the results of the review since it is not dependent on specific features of the included studies. Sensitivity analyses are often conducted by excluding studies that are of poor methodological quality, unpublished, or did not meet all of the inclusion criteria. Sensitivity analyses may also be conducted by reanalyzing data using a range of results from a

---

*Response Models*, 14 STAT. MED. 679, 679-85 (1995); S. J. Smith et al., *On Combining Dose-response Data from Epidemiological Studies by Meta-analysis*, 14 STAT. MED. 531, 531-44 (1995).

[66] Colin B. Begg, *Publication Bias*, *in* THE HANDBOOK OF RESEARCH SYNTHESIS (Harris Cooper and Larry V. Hedges eds. 1994). Parametric tests are used when data are normally distributed, non-parametric tests are used when data are not normally distributed, regression techniques take multiple factors into account, and Bayesian techniques include assessments of prior probabilities.

trial (due to inconsistencies in reporting or how outcomes were measured), reanalyzing data using a range of results for missing data or reanalyzing the data using different statistics.

## CONCLUSION

Systematic reviews and meta-analyses are powerful methods for gathering, critiquing and summarizing medical and scientific information. Meta-analysis is a quantitative approach to systematically combining the results of previous studies.

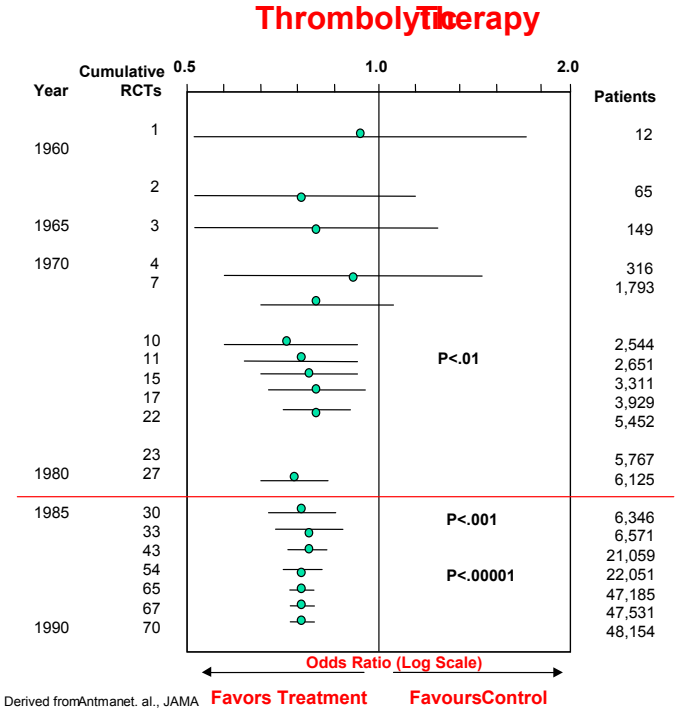
Systematic reviews are combinations of results that adhere to pre-defined methods, but that may not result in quantitative combination of the data. The validity of a systematic review depends on the extent to which the methods of the review reduce random error and systematic bias. Systematic reviews reduce bias because they are conducted according to strictly defined methods that should be pre-specified in a protocol. A good systematic review contains a focused research question, an explicit and comprehensive search strategy, explicit inclusion and exclusion criteria that are uniformly applied by multiple coders, a rigorous critical appraisal of each identified study and, if appropriate, a quantitative summary of the evidence.

Figure 1[67]

[67] The Cochrane Collaboration Home Page, *see supra* note 5.


Figure 2[68]

[68] Antman, *supra* note 9, at 240-48.




Derived fromAntmanet. al., JAMA 1992