Methodological Challenges Associated with Meta-Analyses in Health Care and Behavioral Health Research

Judith A. Shinogle, PhD, MSc
Senior Research Scientist
Maryland Institute for Policy Analysis and Research
University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, Maryland 21250
http://www.umbc.edu/mipar/shinogle.php

Please see http://www.umbc.edu/mipar/shinogle.php for information about the Judith A. Shinogle Memorial Fund, Baltimore, MD, and the AKC Canine Health Foundation (www.akcchf.org), Raleigh, NC. Inquiries may also be directed to Sara Radcliffe, Executive Vice President for Health, Biotechnology Industry Organization, www.bio.org, 202-962-9200.

May 14, 2012

Methodological Challenges Associated with Meta-Analyses in Health Care and Behavioral Health Research

Judith A. Shinogle, PhD, MSc

I. Executive Summary

Meta-analysis is used to inform a wide array of questions ranging from pharmaceutical safety to the relative effectiveness of different medical interventions. Meta-analysis also can be used to generate new hypotheses and reflect on the nature and possible causes of heterogeneity between studies. The wide range of applications has led to an increase in use of meta-analysis. When skillfully conducted, meta-analysis is one way researchers can combine and evaluate different bodies of research to determine the answers to research questions. However, as use of meta-analysis grows, it is imperative that the proper methods are used in order to draw meaningful conclusions from these analyses. This paper provides a framework for determining whether proper methods were used and thus for evaluating the quality of meta-analyses.

Meta-analysis refers to a joint analysis of at least two primary datasets after each has been initially analyzed and published. No new data are produced in a meta-analysis; it is a retrospective study of previous research. The growth of meta-analysis in recent years is due to many factors including an increasing number of systematic reviews to summarize evidence-based medicine, for use in comparative effectiveness studies and the desire to improve the power of clinical trials as well as the precision of studies.

Readers and researchers of meta-analysis should recognize the statistical sophistication needed to combine measures from various data and populations. The methods of meta-analysis should follow the same steps as any well thought out research project – development of research question; systematic review; and collection, evaluation and summarization of data. Each stage should be clearly thought out, and each section well documented. Standards for reporting meta-analysis exist and should be utilized by researchers and readers of studies to evaluate the quality of the study.

A well designed and implemented meta-analysis study can be a great benefit to society. When conducted appropriately, it can reveal the weaknesses and strengths of previous studies, present needed answers to difficult research questions and even provide suggestions for future research. A meta-analysis study can summarize from available studies the effects of interventions across many patients. It can hasten introduction of effective treatments into clinical practice by reducing

false negative results (Type II error). Further, it can determine if the effect of an intervention is sufficiently large in practical, as well as, statistical terms.

While meta-analysis has the potential to be a powerful tool in evaluating health care treatments and interventions, there are many potential pitfalls and problems that are yet to be resolved. A poorly performed meta-analysis can perpetuate biases from ill-conceived studies or lead to false conclusions. This, in turn, can cause consumers and caregivers, who frequently access results of meta-analysis through websites and popular press, to form incorrect conclusions and can result in inappropriate medical decisions. While the idea behind combining studies to improve precision and power is straightforward, the actual implementation of the process is difficult. Those who act or react based on meta-analysis should understand the various biases that could be incorporated into a review. Examples of some potential pitfalls in meta-analysis include publication bias, pipeline bias, and English language bias. In addition combining studies that are not similar in study design, population, methods of analysis or outcome definitions can lead to biases as well, which may result in spurious conclusions being drawn.

II. Introduction

A. Overview of paper

Meta-analysis is a commonly used methodology to combine findings from several studies in order to improve the power of these studies as well as the precision of the estimated effect. This method grew out of the psychology literature and is often used in conjunction with systematic reviews as a means to evaluate health care interventions. This paper will present an overview of meta-analysis technique, provide a description of a sound meta-analysis methodology and discuss challenges and issues involved with developing a meta-analysis study. The paper concludes with a discussion of the use and interpretation of the technique as well as questions that should be asked of a meta-analysis study. Throughout this paper several statistical terms will be used. A table from Koretz (2002) has been recreated in Appendix A and lists many of the key terms and their meanings.

B. Definition of Meta-Analysis

Meta comes from the Greek and means “after.” Meta-analysis refers to a joint analysis of at least two datasets, which maybe from published research or primary data, after each has been initially analyzed and published. The key issue is that meta-analysis is an analysis of analysis as opposed to a re-analysis of primary data. Meta-analysis is an arithmetic tool that is used in conjunction with a systematic review of the literature.

Combining research results dates back to the 1930s with the work of agriculture researchers such as Ronald A. Fisher, Karl Pearson and William G. Cochran. Psychologist Gene Glass reinvented meta-analysis in 1974 to refute claims that psychotherapy was ineffective (Noble, 2006). Since that time the number of meta-

analysis publications has grown dramatically in the psychological literature from a few to over 800 in PsychInfo (Schulze, 2007b). Its use in health care has increased as the call for evidence based medicine and comparative effectiveness increases the need for systematic reviews and summarization of the results.

Meta-analysis is increasingly used to strengthen the results of clinical trials. Many individual clinical trials are underpowered; they contain too few participants to provide a definitive answer to research questions. Researchers attempt to rectify this by combining several trials to improve the statistical power and thus reduce the likelihood of a type II error (failing to determine a difference that truly exists). In addition, by combining results from several trials, meta-analysis has the potential to improve the precision of the estimate of the effect. As the growth of meta-analysis has occurred in health care, it is important to understand what encompasses a sound meta-analysis study as well as the limitations of this technique.

III. Overview of the meta-analysis technique

The methods of meta-analysis should follow the same steps as any well thought out research project. The first step is defining a research question. Next, the researcher samples a defined population of completed studies and then codes relevant methodological characteristics and analyzes the data to test hypothesized questions. Problems will result if each stage is not clearly thought out and each section not well documented.

A. Research Question

A strong meta-analysis starts with a strong, succinct research question. This question should be a simple, single question that is going to be addressed by the study. The researcher should specify the causally related variables and any mediators or interactions that may exist. This question will guide the selection of studies and the rest of the research that is undertaken for the meta-analysis. The goal is to select studies on similar population samples with similar study designs in order to decrease the heterogeneity of studies. To do this, the researcher must carefully identify the pertinent data and studies to ensure an unbiased meta-analysis.

B. Systematic Review

The data collection stage or selection of relevant studies is key to an unbiased meta-analysis study. The selection of the relevant literature for the meta-analysis is done through a systematic review of the literature which is a key component of any meta-analysis (see Appendix B for more detailed discussion of systematic reviews). Some organizations simply do systematic reviews while others incorporate meta-analysis into their systematic review. The review must be performed adequately to not introduce any bias in the analysis. All search strategies utilized in the review must be documented including key terms,

databases searched and limits to search, as well as the number of studies found and the number of studies excluded and reasons for exclusion from the review.

For the systematic review or study selection the researcher should search multiple databases (such as PUBMED, PSYCHINFO, Web of Science, etc.), and it is recommended that the researcher work with a reference librarian to ensure all relevant databases are explored. In addition, the researcher should employ other techniques such as hand searching pertinent journals, searching reference lists of identified articles, personal communications with experts, and clinical trial registries such as http://www.clinicaltrials.gov/.

The goal of the systematic review is to select studies on similar population samples with similar study designs in order to decrease the heterogeneity of studies. All studies must be similar in sampling frame, precise nature of intervention, and measure of success or failure; otherwise, issues of heterogeneity may arise. Heterogeneity in meta-analysis is defined as studies that differ enough to cause concern about combining them (Sutton et al., 2006). The concern with attempting to limit the search to similar studies is that it is rare to find published studies that are identical. If one limits the studies very finely, then the researcher has only one study to review because once a study has been published, most peer reviewed journals do not accept a paper that repeats the same study. However, as the search criteria for studies are relaxed, more differences between the studies occur and hence the heterogeneity of studies increases. Compounding the heterogeneity is that the differences in study design, population, and statistical methods are not always all distinctly included in published papers due to limits on length of papers. This omitted information may bias the conclusions drawn from meta-analysis. On the other hand, too tight a criteria for exclusion of studies could introduce biases such as language, cultural, publication, citation, multiple publication and quality biases discussed below.

Characteristics of what studies to be included and excluded need to be clearly defined at the outset. Examples of issues of study inclusion are: only randomized controlled trials, only English language, and only published studies. While developing specific criteria (and listing them) for including and excluding studies is important, this adds variability. By manipulating these inclusion and exclusion criteria, the researchers may selectively only include studies with positive results and exclude those with negative results or vice versa.

The inclusion and exclusion criteria affect the target population. Many randomized control trials exclude specific patient groups such as those under the age of 18, those with liver or kidney function limitations, or those on specific concomitant medications, for example. If the population samples are so different due to exclusions from the intended target population, this may cause generalizations to be inaccurate as the definition of target population can be vague.

Many times researchers limit their search to studies published in only the English language. Limiting the review to English only language studies may bias results as researchers in other languages tend to publish significant findings that may change

treatment guidelines in English language journals and non-significant or weaker results in non-English language journals.[1]

Along similar lines is the concern that some databases do not include journals published in other countries (especially less developed countries).[2] Studies that are published in journals not indexed in the database used by the researcher results in these studies being unknowingly excluded by the researcher which may add bias to the results. Related biases include pipeline bias which occurs when significant results are published quicker than non-significant results (Sutton et al., 2001).

Meta-analyses often experience publication bias. Most researchers only include literature that has been published in peer reviewed journals. This strategy is often utilized because it is more efficient to search databases of peer reviewed work. In addition, published literature can be challenged and reviewed. Peer review journals do imply a sense of quality, but this quality is only as good as the peer reviewers and the selection process used by the journal. Limiting the review to only published studies can lead to publication bias that can produce Type I error (a study identifying a "significant difference" when the difference is truly due to chance). Studies that show statistically significant differences in results ($p<0.05$) are more likely to have been published than those with non-significant ($p>0.05$) results (Easterbrook et al., 1991).

However, identifying unpublished studies is difficult. Researchers can try to overcome this by examining trial databases and searching the web for "grey" literature in an attempt to secure unpublished studies or non-peer-reviewed studies.[3]

Funnel plot analysis is often used by researchers to examine the studies for publication bias. Funnel plot analysis is a scatterplot of the estimate of effect from each study in the meta-analysis against a measure of its precision (usually 1/[standard error]). In the absence of publication bias the plot resembles a funnel shape as smaller, less precise studies are subject to more random variation than larger studies when estimating an effect. In the presence of publication bias these smaller studies will be absent leading to an asymmetrical funnel plot. The concern with the funnel plot analysis is that it is imprecise and other factors may exist for the asymmetric funnel, such as among study heterogeneity (Peters et al., 2006).

As part of the systematic review, the researcher will often utilize the reference list from collected articles. This may lead to citation bias. Researchers often only cite

[1] For example, a comparison of pairs of articles published by same first author found that 63 percent of trials published in English had produced significant effects compared to 35 percent of trials published in German (Egger et al., 1997).
[2] On the other hand, adding studies from other countries to the meta-analysis could introduce heterogeneity due to difference in culture, medical establishments, etc.
[3] "Grey" literature includes materials that do not enter usual systems of publication, distribution or bibliographic control. Examples of gray literature include technical reports from government agencies or scientific research groups, working papers from research groups or committees, or white papers.

literature that is supportive of their findings which could bias the results of the meta-analysis (Egger et al., 1998).

Another concern in the systematic review strategy is multiple publication bias. If a study results in multiple publications, this too may inflict bias into the meta-analysis results. The inclusion of duplicate data will lead to overestimation of the treatment effects. Multiple publications are more likely to occur from studies with significant results. This may also occur in multicenter trials (Egger et al., 1998).

Study quality in the systematic review is important because including poor or flawed studies may bias the pooled results. Checklists exist to aid in the assessment of study quality (Moher et al. 1995), but many of these checklists are not scientifically created nor have they been validated. Another problem with checklists is that it is impossible to predict which design aspects cause the most bias and it is difficult to predict the direction of the bias. Furthermore, due to space limitations in many journals, the reviewer may not know all the treatment and control conditions for the study. These unobserved or omitted factors of study design could bias the results if they are correlated with any important variables in the study.

The issues above are all threats to validity of a meta-analysis and can be difficult for the reader of a published meta-analysis to detect. The publication bias and variable study quality of included literature are key concerns to the implementation of sound meta-analysis.

C. Collection, Evaluation and Summarization of Data

Data Collection:

Once the systematic review is completed, the next step is data extraction and analysis. This step should be done by at least two people to confirm the analysis. Researchers should have a set plan of data to extract from the study such as estimates of effects, estimates of variations, confounders included in the model, any subgroup analyses, etc. Researchers may use the Q-test (see Appendix A for definition) to detect heterogeneity in studies, but again this test is insensitive. A concern exists that even with a good Q-test those very heterogeneous studies which happen to produce similar numerical data will appear homogeneous. This leads to the “art of meta-analysis”, as it is often better to assess the trials using one’s own judgment. In the data extraction step it is important to confirm the internal consistency of the extractions. The research must be consistent in what is extracted from each study. In addition, the researchers should gather the most information possible regarding data and distribution of the data such as variance of key variables, minimums, maximums, kurtosis or other measures available in the study. Readers of meta-analysis should realize that the review results are dependent on what data are available from the study. Often journals limit the tables and amount of data published which interferes with the quality of the analysis that can be performed.

Another method to combine the data is the Bayesian approach to include the idea of subjective probabilities as opposed to objective probabilities. Before starting the research, the researcher develops a prior belief regarding the outcome. This is often derived from other research in the field. These prior beliefs are combined with the data using Bayes' Theorem. The methods associated with the Bayesian approach are computationally complex, and the development of priors is difficult and again fraught with potential biases. While one of the advantages of the Bayesian approach is that all priors are stated, it is equally important that all prior beliefs should be documented as to where and how they were computed. As more researchers and payment review committees examine and utilize Bayesian approaches, research on the biases and precision associated with these methods should be explored and published.

Evaluation of the Data:

Once the data from the study are collected, the next step is the evaluation of the collected data. In an ideal study, researchers would combine raw data and conduct a pooled analysis. As most researchers do not release primary data, this is often not possible. In fact at this point in time no meta-analysis studies known to the authors have combined the raw data. As the researcher has to combine results from published studies, in general this calls for a combining of reported results such as converting the probability values of statistical significance for each study comparison using a method termed Stouffer's Adding Z's formula (see Appendix C for formula) or converting the relevant estimates of effects into a common metric. Many concerns exist about combining the data as the researcher is limited to what data are presented in the published study, which often excludes higher moments of the data distribution (such as the distribution of the data or the shape of the data in the tails).

If the outcome is expressed as a dichotomous variable, the summary estimate can be expressed as: 1) absolute risk difference; 2) relative risk ratio; or 3) odds ratio. If the outcome is expressed as a continuous variable then the summary estimate is usually expressed as the weighted mean difference (where weights are developed based on study quality). If the continuous variable is not reported in the same units for each study in the review, the summary estimate may be expressed as the standardized mean difference. Often, meta-analysis uses confidence intervals instead of *p*-values in presentation of the distribution around the effect.

Combining the data from the studies reviewed allows the researcher to estimate global means or perform meta-regression. Yet, many concerns exist in combining the data from divergent studies. Firstly, difference in rules for inference in the primary studies can lead the meta-analysis to suggest a causal effect where none exists (*i.e.*, weighting equally experimental, quasi-experimental and case studies) or weighting equally studies with different sample sizes. This is especially a concern when the content from primary studies provides too little detail about how latent constructs and theories are operationalized which can mask the influence of interacting or mediating variables that serve to explain the cause and effects relationships. Secondly, attrition rates or missing values may be interpreted by the

original researcher differently than by the reviewer. Some studies may view attrition due to success of the intervention, while others may view attrition as due to failure of the intervention. Thirdly, potential heterogeneity can be added when studies develop a concept for their outcome, but each study may utilize a different measure for the concept. Finally, another form of adding heterogeneity in studies that is often not explored is this concept of measurement correspondence. Studies may have a similar underlying concept they are measuring, such as levels of depression, but utilize different measurements to examine this outcome.

There are several ways to address the heterogeneity in a meta-analysis study. One method is to utilize a random effects analysis (see Appendix A); another is to perform a sub-analysis by variable that may lead to the heterogeneity of results. This requires adequate study size, as well as comparable subpopulation data from the literature which is often not available or is not interpretable.

Scoring systems can be used to rate the quality of the study; however, none of the proposed systems have been validated or widely accepted. These same scoring systems can impose a cut-off value of studies to include in the analysis. Another method to summarize the data that attempts to quantify quality is to use the quality scale to weight study results or incorporate the score in the standard precision weighting. Weighting in this fashion may add another level of subjectivity into the meta-analysis study design.

The researcher can use meta-regression which incorporates a quality score or individual markers (such as degree of blinding, method of treatment allocation, etc.) in a regression model as explanatory variables. Several issues exist regarding the interpretation and the reliability of meta-regression (Thompson and Higgins, 2005). A report conducted by Agency for Healthcare Research & Quality (AHRQ) found several issues regarding meta-regression (Moton et al., 2004). First, failure to include important covariates at any level can bias the results. Even with this issue, a model that includes a covariate that is an aggregate of a person-level characteristic (*i.e.*, percent of females), rather than a study characteristic (randomization), can produce biased results. The trade-off between the biases of incorporating an aggregated covariate versus excluding it requires further exploration (Morton et al., 2004). Meta-regression may also be utilized to examine the determinants of the effects size. More research is required to fully understand the statistical properties and potential biases that could occur in this method.

Summarization of the Data:

Several statistical issues exist regarding the summarization and analysis utilized in meta-analysis that is yet to be resolved. It is problematic when the summary estimates from the literature are presented as odds ratios as there is a correlation between log of odds ratio and its standard error. These summary estimates are developed from data with non-normal distributions and thus the simple aggregation may or may not be appropriate. Egger's regression test (effect/standard error is regressed on a measure of precision, generally 1/standard error) faces the same issue as funnel plots, as the effects and standard error are correlated. An

alternative is to use a simple weighted linear regression with log odds ratio as the dependent variable and the inverse of the total sample size as the independent variable (Peters et al., 2006). Peters et al. (2006) examined each of these measures and found that neither the Eggers regression nor the alternative is particularly powerful in the various scenarios analyzed.

When summarizing, data researchers should be cognizant of what is in the study as well as what information may be missing. For example, at the evaluation stage, missing information such as on adverse events can lead to unreliable conclusions. Users of meta-analysis should recognize how the researcher summarized data and any potential issues with the outcomes that could affect the combining of data.

D. Standards in Reporting

In an attempt to improve the reporting of meta-analysis studies (not necessarily the quality of the studies), guidelines were developed for the reporting of meta-analysis of randomized controlled trials by the Quality of Reporting of Meta-analyses (QUORUM) group. These guidelines grew out of a conference of the QUORUM group which consisted of about 30 members including clinical epidemiologists, clinicians, statisticians, journal editors and researchers. See Table 1 for detail of the reporting required in many medical journals.

Table 1. Quality of Reporting of Meta-analysis (QUORUM)

Heading	Subheading	Descriptor
Title		Identify the report as a meta-analysis (or systematic review) of RCTs
Abstract		Use a structured format
	Objectives	Describes the clinical question explicitly
	Data sources	The databases (*i.e.*, lists) and other information sources
	Review methods	The selection criteria (*i.e.*, population, intervention, outcome, and study design); methods for validity assessment, data abstraction, and study characteristics, and quantitative data synthesis in sufficient detail to permit replication
	Results	Characteristics of the RCTs included and excluded; quantitative and qualitative findings (point estimates and confidence intervals) and subgroup analysis
	Conclusion	The main results
Introduction		The explicit clinical problem, biological rationale for the intervention and rationale for review
Methods	Searching	The information sources in detail (databases, registrars, personal files, expert informants, hand

		searching) and any restrictions (years, publication status, language of publication)
	Selection	The inclusion and exclusion criteria (defining population, intervention, principal outcomes and study design). Include numbers in a flow chart
	Validity assessment	The criteria and process used (*e.g.*, masked conditions, quality assessment, and their findings
	Data abstraction	The process or processes used (*e.g.*, completed independently, in duplication)
	Study characteristics	The type of study design, participants' characteristics, details of intervention, outcome definitions and how clinical heterogeneity was assessed.
	Quantitative data synthesis	The principal measures of effect (*e.g.*, relative risk), method of combining results (statistical testing and confidence intervals), handling of missing data, how statistical heterogeneity was assessed, rational for any a-priori sensitivity analysis and subgroup analysis and any assessment of publication bias
Results	Trial flow	Provide a meta-analysis profile summarizing trial flow in a diagram
	Study characteristics	Present descriptive data for each trial (*e.g.*, age, sample size, intervention, dose, duration, follow up period)
	Quantitative data synthesis	Report agreement on the selection and validity assessment; present simple summary results (for each treatment group in each trial, for each primary outcome); present data needed to calculate effect sizes and confidence intervals in intention to treat analysis (*e.g.*, tables of counts, means, and standard deviations, proportions)
Discussion		Summarize key findings; discuss clinical inferences based on internal and external validity; interpret the results in light of the totality of available evidence; describe potential biases in the review process (*e.g.*, publication bias); and suggest a future research agenda

Source: Moher et al., 1999.

This standardization in reporting of meta-analysis is a good first step at improving the quality of meta-analysis but does not necessarily improve the quality unless journals require submission of all components by the author and provide all components to the reviewer.

These recommendations only address meta-analysis of randomized control trials, which leaves out some potentially important sources for meta-analyses from non-

experimental and quasi-experimental studies (more real-world study designs.) While excluding these may improve the homogeneity of studies, it does limit the breadth of the reviews.

However, including non-randomized control studies adds another level of complexity in synthesis. To address this issue the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Working Group developed a proposal for reporting (Stroup et al., 2000). Table 2 lists a reporting checklist. Even if authors of meta-analyses follow these recommendations and adhere to transparency on the systematic review and summarization of data, biases may still exist that are not known to the reader. In addition, the reviewers of these meta-analysis articles must know and adhere to these guidelines.

Table 2. Proposed Reporting Checklist for Meta-Analysis of Observational Studies	
Reporting of background should include:	
	Problem definition
	Hypothesis statement
	Description of study outcome(s)
	Type of exposure or intervention used
	Type of study design used
	Study population
Reporting of search strategy should include:	
	Qualifications of searchers (*e.g.*, librarian and investigators)
	Search strategy including time period included in the synthesis and keywords
	Effort to include all available studies, including contact with authors
	Databases and registries searched
	Search software used, name and version, including special features used (*e.g.*, explosion)
	Use of hand searching (*e.g.*, reference lists of obtained articles)
	List of citations located and those excluded, including justification
	Method of addressing articles published in languages other than English
	Method of handling abstracts and unpublished studies
	Description of any contact with the authors
Reporting of methods should include:	
	Description of relevance or appropriateness of studies assembled for assessing the hypothesis to be tested
	Rationale for selection and coding of data (*e.g.*, sound clinical principles or convenience)
	Documentation of how data were classified and coded (*e.g.*, multiple raters,

	blinding, and interrater reliability
	Assessment of confounding (*e.g.*, comparability of cases and controls in studies where appropriate)
	Assessment of study quality including blinding of quality assessors; stratification or regression on possible predictors of study results
	Assessment of heterogeneity
	Description of statistical methods (*e.g.*, complete description of fixed or random effects models, justification of whether the chosen models account for predictors of study results, dose-response models, or cumulative meta-analysis) in sufficient detail to replicate
	Provision of appropriate tables and graphics
Reporting of results should include:	
	Graphic summarizing individual study estimates and overall estimate
	Table giving descriptive information for each study included
	Results of sensitivity testing (*e.g.*, subgroup analysis)
	Indication of statistical uncertainty of findings
Reporting of discussion should include:	
	Quantitative assessment of bias (*e.g.*, publication bias)
	Justification for exclusion (*e.g.*, exclusion of non-English language citations)
	Assessment of quality of included studies
Reporting of conclusions should include:	
	Consideration of alternative explanations for observed results
	Generalization of the conclusions (*i.e.*, appropriate for the data presented and within the domain of the literature review)
	Guidelines for future research
	Disclosure of funding source

Source: Stroup et al., 2000

From the tables above, a sound meta-analysis requires a significant amount of documentation and reporting. However, if the researcher conducts a sound meta-analysis, journal length limitations may not allow adequate reporting. This leaves the quality of the meta-analysis in question to the reader and also leaves doubt whether such analyses should impact decisions related to policy and the public health.

IV. Conclusions

Meta-analysis is a mathematical operation; the thinking process is the systematic review. When conducted appropriately, meta-analysis can provide valuable information to policymakers, clinicians, patients and others on the overall effects of

treatments and programs and generate questions for future research. Meta-analysis can be used to generate new hypotheses and reflect on the nature and possible causes of heterogeneity between studies. Done correctly, meta-analysis can provide information to guide in clinicians' and patients' health care decision making.

While meta-analysis has the potential to be a powerful tool in evaluating health care treatments (see Appendix D for list of some strengths of meta-analysis summarized from Nobel JH, 2006) and interventions, there are many potential pitfalls and problems that are yet to be resolved (See Appendix E for weaknesses, again summarized from Nobel JH, 2006). If done inappropriately, it could lead to inappropriate conclusions. Researchers and policymakers need to recognize that meta-analysis produces no new data; the analysis by definition is retrospective. Further, the synthesized data rarely allows causal inferences unless all the underlying studies in the review allow for causal interpretations. While conceptually, the idea behind combining studies to improve precision and power appears important in developing evidence-based practice, the actual implementation of the process is difficult.

Results or conclusions drawn from a meta-analysis can inform but should not direct decision makers when extensive heterogeneity in study participants, outcomes, and methods are included in meta-analytic research. While many use meta-analyses as new research and often interpret the results as causal, some caution should exist in these interpretations, especially if the study designs are weak or documentation not available for review.

As one reviews a meta-analysis, there are several questions the reader should ask:

- Does the paper have a well defined research question?
- Does the research have a well defined target population?
- Does the research have explicit inclusion and exclusion criteria that fit both the first two questions?
- Does the research list the search strategy, what databases were searched, results from searches, the number of studies that were excluded and why they were excluded?
- In combining the studies, does the researcher account for varying quality of studies?
- In combining the studies, does the researcher examine all potential data metrics (mean, variance and other measures of the distribution)?
- Do the sample and data in the combined studies fit the target population and address the research question?
- Does the researcher test for heterogeneity?

- Does the researcher provide a sensitivity analysis?
- What information is missing in the analysis?
- Are the results from the meta-analysis in line with other results?

Meta-analysis, when conducted appropriately, can reveal the weaknesses and strengths of previous studies as well as point to future studies and provide useful, important information. Thus, meta-analysis can provide a valid method to examine various studies. On the other hand, poorly performed meta-analysis can perpetuate biases from ill-conceived studies or lead to false conclusions if the study is not performed according to protocol. Researchers should address the potential biases when conducting meta-analysis and systematic reviews. Several working groups have put together checklists, which meta-analysis researchers should reference when conducting their analyses, and which readers should consult to help determine the quality of the study. These checklists need to be evaluated and refined as newer methods are utilized such as Bayesian methods. Even with the check lists as an aid, it can be difficult for the reader of a meta-analysis study to determine the quality of the study without adequate reporting. Thus, the conclusions of a meta-analysis study must be viewed cautiously. As researchers, including government entities, expand the utilization of meta-analysis, a need exists to improve the science. Future research should develop standards for all meta-analysis techniques and provide adequate means for reporting and critiquing studies.

May 14, 2012

References:

Campbell D, Stanley J. 1963. Experimental and Quasi-experimental Designs for Research. Chicago, IL. Rand McNally.

Copper H. 1997. The Integrative Research Review. Beverley Hills, CA. Sage. 1997:15.

Cooper H, Hedges LV. 1994. Potentials and Limitations of Research Synthesis. In Cooper H, Hedges LV. Eds. The Handbook of Research Synthesis. New York. Russel Sage.

Durlak JA. 2003. Basic Principles of Meta-Analysis. In Handbook of Research Methods in Clinical Psychology. Robers MC, Ilardi SS eds. Blackwell Publishing. Malden MA. USA.

Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. 1991. Publication bias in Clinical Research. *Lancet.* 336:887-872.

Egger M, Smith GD. 1998. Meta-analysis: Bias in Location and Selection of Studies. *British Medical Journal*. 316:51-66.

Egger M, Smith GD. 1997. Meta-analysis: Potentials and Promise. *British Medical Journal.* 315:1371-1374.

Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. 1997. Language Bias in Randomized Control Trials published in English and German. *Lancet.* 350:326-329.

Koretz. RL. 2002. Methods of Meta-Analysis: An Analysis. *Current Opinion in Clinical Nutrition and Metabolic Care.* 5(5):467-474.

Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF, for the QUORUM Group. 1999. Improving the Quality of Reports of Meta-analysis of Randomized Control Trials: The QUORUM Statement. *The Lancet.* 354:1896-1901.

Moncrieff J. 1998. Research Synthesis: Systematic Reviews and Meta-analyses. *International Review of Psychiatry.* 10:304-311.

Morton SC, Adams JL, Suttorp MJ, Shekelle PG. 2004 Meta-regression Approaches: What, Why, When, and How? Technical Review 8 (Prepared by Southern California-RAND Evidence-based Practice Center, under Contract No 290-97-0001). AHRQ

Publication No. 04-0033. Rockville, MD: Agency for Healthcare Research and Quality. March.

Noble JH. 1974. Peer Review: Quality Control of Applied Social Research. *Science.* 185:916-921.

Noble JH. 2006. Meta-analysis: Methods, strengths, weaknesses and political issues. *The Journal of Laboratory and Clinical Medicine.* 147(1):7-20.

Peters JR, Sutton AJ, Jones DR, Abrams KR, Rushton L. 2006. Comparison of Two Methods to Detect Publication Bias in Meta-analysis. *JAMA.* 295 (6):676-680.

Schulze R. 2007a. Current Methods for Meta-Analysis: Approaches, Issues and Developments. *Journal of Psychology.* 215(2):90-103.

Schulze R. 2007b. The State and the Art of Meta-Analysis. *Journal of Psychology.* 215(2):87-89.

Smith GD, Egger M. 1998. Meta-analysis: Unresolved Issues and Future Developments. *British Medical Journal.* 316:221-225.

Stroup DF, Berlin JA Morton SC, Olin I, Wiliamson GD, Rennie D, Moher D, Sipe TA, Thacker SB, for the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Group. 2000. Meta-analysis of Observational Studies in Epidemiology. *JAMA.* 283(13):2008-2012 (April 19, 2000).

Sutton AJ, Abram KR, Jones DR. 2001. An Illustrated Guide to the Methods of Meta-Analysis. *Journal of Evaluation in Clinical Practice.* 7(2):135-148.

Thompson SG, Higgins JPT. 2005. Treating Individuals 4: Can Meta-analysis Help Target Interventions at Individuals Most Likely to Benefit? *Lancet.* 365:341-346.

Webb E, Campbell D, Schwartz R, Sechrest L, Grove J. 1981. Nonreactive Measures in the Social Sciences. Boston, MA. Houghton Mifflin.

Wolf FM. 1986. Meta-analysis: Quantitative Methods for Research Synthesis. Beverly Hills, CA. Sage.

APPENDIX A – DEFINITION OF TERMS

Term	Meaning
Absolute risk difference	The arithmetic difference between the rate of the event in question in the treated group and the rate of that event in the control group
Confidence interval	If the experiment were to be repeated a hundred times, the 'n'% (typically 95%) confidence interval quantifies the limits between which the observed result would fall 'n' of those times. (In other words, the n% confidence interval means that there is an n% likelihood that the true answer lies somewhere with that range.) In general, a meaningful difference is present if the 95% confidence interval does not overlap the line of no effect (0.0 for absolute risk difference or weighted/standardized mean differences, 1.0 for relative risks or odds ratios).
Continuous variable	A result that is not discrete (dichotomous).
Data combination	The act of adding together observed results from two or more studies.
Data pooling	Data combination employing the simple addition of all results without any weights.
Dichotomous variable	A result that is either present or absent. (Categorical variables are results that can be classified as belonging to one or a limited number of possibilities; if that limited number is two, categorical and dichotomous have identical meanings.)
Fixed effects model	The mathematical model for meta-analysis that is used if the trials are believed to be homogeneous. (The model only employs intra-trial variances in the weighting system.)
Forest plot	The pictorial presentation of meta-analysis
Funnel plot	A technique to detect publication bias. (The size of the effect is graphed against the size of the study. If no publication bias is present, the small trials should be equally distributed on both sides of the largest one or two studies [forming a bell shaped curve or funnel].)
Heterogeneity	Studies that are unlike enough to cause concern about the appropriateness of employing data combination (or at least to result in modifications of the analytic procedures that are employed.)
Homogeneity	Studies are alike enough to permit data combination
Line of no effect	Vertical line in Forest plot that represents equivalency in outcome for the two treatment arms (1.0 for relative risk ratio or odds ratio; 0.0 for absolute risk difference or weighted/standardized mean difference).
Meta-analysis	Data combination employing a weighting process for the individual

	studies such that those with less variance are given more weight.
Narrative review	A putative expert's overview of a topic in which no formal rules or standards need be used in its preparation.
Number needed to treat	The number of patients that would have to be given the intervention under study in order to achieve or avoid one event (calculated by dividing 100 by the absolute risk difference expressed as a percentage). (If there is a 5% difference favoring the treatment group, one would need to treat 20 patients to obtain one favorable outcome.)
Odds ratio	The odds of a particular event occurring in the intervention group divided by the odds of it occurring in the control group. (The odds ratio is not easy to grasp intuitively; when the rates of the *event are small in each group*, the odds ratio approximates the relative risk ratio).
Power	The likelihood that a particular difference will be statistically demonstrated given the number of individuals enrolled in a trial and the anticipated event rates in the interventional and control groups.
Precision	An expression of how close the observed result is to the true situation. (A high precision is reflected by a narrow confidence interval.)
Publication bias	The tendency for dramatic or significant results to be submitted to, and accepted by, medical journals.
Q-test	A statistical estimate of the presence of homogeneity.
Quality of the study	An evaluation of the scientific rigor (elements that reduce the influence of bias) of the study. (High quality indicates the presence of good rigor.)
Quality score	A numerical estimate of the quality of the study.
Quasi-randomized	A method of allocation whereby the investigator or subject will know into which study arm assignment will occur prior to the subject being enrolled (*e.g.*, assignment made by patient identification number or day of week).
Random effects model	The mathematical model for meta-analysis that is used if the trials are believed to be heterogeneous. (The model employs both intra-trial and inter-trial variances in the weighting system.)
Relative Risk Ratio	The incidence of the event in the treated group divided by the incidence of the event in the control group. (For example, a risk ratio of 0.75 means that the event occurred only ¾ as often in the treated group as in the control group.)
Review article	See Narrative review, Systematic review
Sensitivity analysis	Confining the meta-analysis to a subgroup of trials that share a

	certain feature. (An a priori decision is made regarding what the important components of heterogeneity are likely to be and separate analysis is performed that only include trials that share these components.)
Standardized mean difference	The meta-analysis summary estimate for studies in which the outcome variable is continuous but not comparable between studies (*e.g.*, expressing the results of test as a percentage of the upper limit of normal.)
Subgroup analysis	See Sensitivity analysis.
Summary estimate	The result of the meta-analytic calculation.
Systematic review	A structured effort to obtain the answer to a specific question by employing preplanned literature searches, data abstraction, and analyses.
Type I error	A study describing 'significant difference' in which the observed difference is truly due to chance.
Type II error	A study failing to see a difference that truly exists.
Variable	See Continuous variable; Dichotomous variable.
Variance	A statistical estimate indicating how far a particular observation is from the truth. (It reflects how close the individual observations are to each other, using the standard deviation or confidence interval.) (Although other factors play into the calculation of variance, larger trials usually have less variance.)
Weighted mean difference	The meta-analysis summary estimate for studies in which the outcome variable is continuous.
Weighting	The degree of emphasis that will be given to each study in the meta-analysis. (It is inversely proportional to the variance.)

Source: Koretz, 2002

APPENDIX B- SYSTEMATIC REVIEWS

A systematic review is a structured review that attempts to apply scientific strategies to limit bias to the synthesis of all relevant studies that address a specific research question. Because the review process could be subject to bias, a useful review requires clear reporting of information obtained using rigorous methods. Systematic reviews are scientific investigations and thus should include pre-planned methods and an assembly of original studies as their "data." The review then synthesizes the results of multiple other studies by using strategies that limit bias and random error. These strategies include a comprehensive search of all potentially relevant articles and the use of explicit, reproducible criteria in the selection of articles for review. Primary research designs and study characteristics are appraised, data are synthesized, and results are interpreted.

Systematic reviews are generated to answer specific, often narrow, research questions. These questions should be formulated explicitly according to four components: a specific population and setting; the condition of interest; an exposure to a test, treatment or intervention; and one or more specific outcomes. If the question that is driving the review is not clear from the title, abstract, or introduction, or if no methods section is included, the paper is more likely to be a narrative review than a systematic review.

APPENDIX C – STOUFFERS'S ADDING Z FORMULA

Stouffer's Adding Zs formula:

$$Z_w = \frac{\sum_{1=1}^{N} W_i Z_i}{\sqrt{\sum_{i=1}^{N} W_i^2}}$$

Z_w = the standard normal deviate or Z-score for weighted combination of studies
Z_i = the standard normal deviate (or Z-score) for the i^{th} comparison
W_i = the weighting factor for each study
N = total number of studies

APPENDIX D – STRENGTHS OF META-ANALYSIS

Strength	**Source**
Can summarize from available studies the effects of interventions across many patients.	Thomson and Higgins, 2005
Can reveal research designs as moderators of study results.	Cooper, 1997
Can determine if the effect of the intervention is sufficiently large in practical as well as statistical terms.	Lipsey and Wilson, 1993
Can, through multi-operationalism, reduce uncertainty of interpretation – if the measures encompassed are sufficiently valid.	Webb et al., 1981
Can allow more objective assessment of evidence and thereby reduce disagreement.	Egger and Smith, 1997
Can reduce false negative results and thereby hasten introduction of effective treatments into clinical practice.	Egger and Smith, 1997
Can allow testing a priori hypotheses about treatment effects in subgroups.	Egger and Smith, 1997
Can clarify heterogeneity between study results.	Egger and Smith, 1997
Can suggest promising questions for future study.	Egger and Smith, 1997
Can assist accurate calculation of sample size needed for future studies.	Egger and Smith, 1997
Can increase precision of literature reviews.	Cooper and Hedges, 1994
Can, through consistent coding of primary study characteristics and use of multiple judges, reduce bias in judgments about the “quality” of individual studies.	Cooper and Hedges, 1994
Can, through various statistical formulas, provide confidence interval calculation of effect size estimates.	Cooper and Hedges, 1994
Can, using different assumptions or alternative statistical models, clarify and interpret the range of possible conclusions about the “quality” of segments of the literature review.	Cooper and Hedges, 1994
Can overcome problems of traditional literature reviews involving: (a) selective inclusion of studies, (b) subjective	Wolf, 1986

weighting of studies and their interpretation, (c) failure to examine study characteristics as source of disparate or consistent results across studies, and (d) failure to address influences of moderating variables in the relationships examined.	
Can, through systematic use of threats-to-inference framework, reveal structural flaws and sources of bias in research procedures.	Campbell and Stanley, 1962; Nobel, 1974

Source: Noble JH, 2006

APPENDIX E – WEAKNESSES OF META-ANALYSIS

Weakness	**Source**
Can pass along inflated estimates of size effects based on: (a) research design characteristics, (b) published vs. unpublished study composition, and (c) small sample sizes.	Lipsey and Wilson, 1993
Can be limited by unspecified “black box” treatment and control conditions as well as lack of attention to potential interactions with subject characteristics, range of outcomes and temporal factors.	Lipsey and Wilson, 1993
Can be compromised by inclusion of non-peer reviewed data – especially if derived from biased sources.	Smith and Egger, 1998
Application of results of meta-analysis to individual patients remains a difficult judgment call because “uncertainty with respect to a particular patient will always be greater than with respect to overall patient group.”	Smith and Egger, 1998
Interpretation of different attrition rates as indicative of therapeutic success or failure is problematic because of judgment subjectivity.	Smith and Egger, 1998
Cannot overcome subjectivity in choice of outcomes and their weighting in analysis.	Smith and Egger, 1998
The combined statistically significant estimate of size effect may still prove inconclusive because of ignorance or uncertainty about other relevant matters.	Smith and Egger, 1998
The correlation nature of review-generated evidence precludes conclusive inference about the strength of possible confounders or rival explanations for the reported effects of primary studies.	Cooper and Hedges, 1994
The post hoc nature of research synthesis prevents use of review-generated evidence to develop and test theory simultaneously.	Cooper and Hedges, 1994
Partially complete information in some studies compromises coding certainty and reliability as well as the extent to which studies with complete information can represent the universe of all relevant studies.	Cooper and Hedges, 1994
Reader judgments about the “quality” of a specified research synthesis are dependent on such subjective criteria as: (a) self-assessed ability to interpret the review, (b) organization, (c) writing style, (d) clarity of	Cooper and Hedges, 1994

focus, (e) use of citations, (f) attention to variable definition, (g) attention to details of methodology, and (h) manuscript layout.	
Cannot eliminate, without registration of all trials, publication bias wherein positive findings get published and negative ones do not.	Moncrieff, 1998
Synthesis may mask relevant heterogeneity with respect to distinguishing characteristics of participants, circumstances of interventions, and the conduct of primary research.	Moncrieff, 1998
Arithmetic nature of meta-analysis can produce false impression of certainty in an inherently uncertain process with many subjective elements	Moncrieff, 1998
Statistics for calculating “heterogeneity” in primary study size effects are weak and lack power to distinguish true differences among trials from chance effects; hence, statistical non-significance does not mean there is sufficient homogeneity to justify their combination.	Moncrieff, 1998; Thompson and Higgins, 2005
Meta-regression, in which treatment benefit is related, where possible, to some average characteristic of patients in each trial mean age or proportion of women), is fraught with interpretation difficulty and is generally misleading.	Thompson and Higgins, 2005
Meta-regression describes observed relationships across trials that – even if all randomized – are still subject to confounding by other variables that may vary between trials and thus invalidate inferences about relationship cause and effect.	Thompson and Higgins, 2005

Source: Noble JH, 2006