Statistics Commentary Series

Commentary No. 18: Meta-Analysis, Part 1 – What It Is

David L. Streiner, PhD, CPsych

If you had used Medline before 1980 to do a search for meta-analyses (MAs), you would have come up empty handed. Difficult as it may be to believe, the term was introduced only in 1977,[1] and did not appear in Medline until the early 1980s. When I was writing this article, I used the term in a search and came up with over 78,000 hits. This exponential growth in its use raises 2 important questions: (1) what is MA, and (2) what should you, as a doer or consumer of them, be wary of? In this, the first of a 2-part series, I will address the first issue; the next article will examine the second.

Very briefly, MA is a method for combining the results of many studies in order to arrive at a better estimate of "truth," which usually is about the effectiveness of some intervention, although it can also be used to answer questions about diagnostic test accuracy and the natural history of a disorder. Cynics would define it somewhat differently, as a method of combining the results of many inadequate studies to arrive at a single, erroneous answer. Not surprisingly, the reality is somewhere between these 2 extreme positions. The need, though, for MA is not disputed. Prior to its advent, combining the results of many studies was done in 1 of 2 ways. The traditional method, still found in annual reviews published in many specialties, relied on an expert – self-described or otherwise – to summarize the research in some area. This raises the obvious question of any biases that the reviewer may have; after all, who would spend the time to read all of the articles and summarize their findings unless he or she had some involvement in the field? The biases could involve selecting only those articles that supported the reviewer's position, or being far more critical of articles with contrary findings than those with supporting conclusions. Even if the reviewer wasn't biased, there could be questions raised about the completeness of the reviewer's search.

The second method of summarizing the results of many studies is called vote counting, and the name is a perfect description of how it's done: gather up all the articles you can find about a topic and simply count how many had positive outcomes and how many had negative ones. However, in addition to the same issue of the completeness of the search, there are 2 other problems. The first is that there may be a tie. For example, 1 article looked at the relationship between socioeconomic status and obesity among American males.[2] They found 27 articles: 12 concluded that there was a positive correlation, 12 a negative one, and 3 said there was no association. The second problem is a more serious one; a study that was poorly executed and with a small sample size is given the same weight as one that was well done and had a very large sample size.

The term "meta-analysis" was coined by Gene Glass[3] and first applied in a summary of articles about the effectiveness of psychotherapy,[1] although the statistician, Karl Pearson, had used a similar approach to combine many studies of the effectiveness of inoculation against typhoid fever many years earlier.[4] Since that inaugural article, there have been a number of improvements in how MAs are conducted, but the basic elements have remained the same.

The first step, which has been greatly facilitated by on-line databases such as Medline, PsycInfo, Embase, CINAHL, and dozens of others, consists of a thorough search for all articles that meet certain criteria, such as the population (eg, age limits, the disorder of interest) and the methodology (eg, only randomized controlled trials), although they may include some secondary criteria such as the language of publication. This is often supplemented by hand searching key journals and checking the reference lists of articles, as well as communicating with other researchers to identify any studies that may have been missed. The second step is to review the identified articles, because most of those flagged by the databases will not be relevant. Ideally, this is done independently by 2 reviewers, using a checklist. At this point, other criteria are often specified, such as a minimum sample size of the study, a minimum limit on the length of treatment or follow-up, how the diagnosis was established (eg, clinical judgment versus a structured interview), the outcome measure, and so forth. The important point is that the outcome itself – positive or negative – should never be used as a criterion. In the next article, I will discuss how this step can lead various MAs in the same area to come to different conclusions.

From the Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton; and Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada.
Reprints: David L. Streiner, PhD, CPsych, St Joseph's Healthcare, Mountain Campus, 100 W 5th St, Hamilton, Ontario, Canada L8N 3K7 (e-mail: streiner@mcmaster.ca).

ISSN: 0271-0749
DOI: 10.1097/JCP.0000000000000625

The articles that pass muster are then abstracted and the relevant information entered into a database or a program designed for MA, such as RevMan,[5] which was developed by the Cochrane Collaboration and has the distinct advantage that it is free. Again, this should ideally be done by 2 independent abstractors in order to check the reliability of the process, and differences resolved either through discussion or by a third party. All this work yields a single number for each outcome: an effect size (ES), which is most usually the standardized mean difference for continuous outcomes, or an odds ratio for dichotomous ones.[6]

What has been done so far is a *systematic review*, in which the conclusions can be summarized in narrative form by the author. What differentiates a systematic review from a MA is what comes next – mathematically combining the ESs to arrive at a global estimate of the magnitude of the effect. But there are some steps that need to be taken beforehand. First, the global ES is not simply the mean of the individual ESs. As I mentioned while discussing the vote counting approach, there is a problem if all studies are considered equal, irrespective of their sample sizes. To avoid this problem, each study's ES is multiplied by some value that is a function of the sample size, which is usually the reciprocal of the variance. Second, an attempt is made to determine the presence and extent of the "file drawer problem."[7] This is based on the fact that studies with negative findings are less likely to be submitted than those with positive findings,[8] and if they are submitted, they are less likely to be published.[9] Consequently, what finds its way into journals may be only a sample of all studies, biased toward positive results because those with negative findings are languishing, unloved and unpublished, in researchers' filing cabinets. A number of visual and analytic methods exist to try to estimate the magnitude of this publication bias, including funnel plots, Egger's test, and the trim and fill method.[10] However, these techniques work only for larger accumulations of articles and are unreliable if there are fewer than 10 or 15 of them. The last pre-calculation step is to look at the degree of heterogeneity among the studies' ESs, which is usually done with 2 tests. The Q statistic determines if it is significant, but suffers from the same problem as all chi-squared based statistics: it is insensitive if the number of studies is too small, and is too sensitive to heterogeneity if there are too many (and nobody knows what the middle range is). The I^2 statistic quantifies the degree of homogeneity, ranging from 0% to 100%, and the general rule of thumb is that 50% or higher reflects a significant (albeit not in the statistical sense) amount. There is a heated debate regarding what to do if there is heterogeneity – eliminate outlying articles until the resulting ESs are homogeneous, or live with it and try to determine its causes. This will be discussed further in the next article.

After all this preliminary work, the last step is relatively easy: determining the mean weighted ES. This is a parameter, and as with all parameters, it is reported with its accompanying confidence interval (CI).[6] If the ES is reported as a standardized mean difference, then the 95% CI should not include zero for the result to be statistically significant; if it is reported as an odds ratio, then it should not include 1.

Although this is the last required step, there is actually a further analysis that can be done, called "meta-regression." This sounds formidable, but is actually nothing more than a regular multiple regression. The dependent variable is the ES for each study, and the predictor variables are factors that could have played a role in determining its size; things like an evaluation of how well the study was done, the mean medication dosage or the number of therapy sessions, the length of the follow-up, the sample size, and so forth. For example, we showed that the magnitude of the ES for the efficacy of tricyclic antidepressants was affected by how the diagnosis was made – larger when it was based on objective criteria than on the psychiatrists' judgment, presumably because some patients in the latter group were not truly depressed.[11] One non-mathematical variant of a sub-analysis of a MA is a "cumulative MA," in which the articles are arranged chronologically and the MA is done after the publication date for each article. One such study showed that the effectiveness of treatments for a myocardial infarction (MI) was demonstrated by 1970.[12] This had two implications. First, textbooks didn't mention this until 1980, showing that reliance on them for up-to-date information about treatment is fraught with danger; and second, that the thousands of patients who were entered into trials after 1970 never should have been, because the issue had already been settled.

To sum up, MA is a very powerful technique for teasing truth out of conflicting or under-powered studies. As an example, there were a number of randomized controlled trials of aspirin to reduce the recurrence of MIs. All were suggestive but none was statistically significant, likely because they did not have sufficient power to detect a dichotomous outcome for a rare event. An MA combining all of these under-powered studies, though, showed conclusively that aspirin is effective,[13] so that now it would be inconceivable that an at-risk person would not be told to use it. However, MA does not always provide unequivocal answers. There are many instances of MAs coming to opposite conclusions. In the next article, I will discuss some of the difficulties in conducting MAs and why they can yield conflicting results.

AUTHOR DISCLOSURE INFORMATION

The author declares no conflicts of interest.

REFERENCES

1. Smith ML, Glass GV. Meta-analysis of psychotherapy outcome studies. *Am Psychol*. 1977;32:752–760.
2. Sobal J, Stunkard AJ. Socioeconomic status and obesity: a review of the literature. *Psychol Bull*. 1989;105:260–275.
3. Glass GV. Primary, secondary and meta-analysis of research. *Educ Researcher*. 1976;10:3–8.
4. Pearson K. Report on certain enteric fever inoculation statistics. *BMJ*. 1904; 3:1243–1246.
5. Cochrane Informatics & Knowledge Management Department. RevMan 5 download and installation. Available at: http://tech.cochrane.org/revman/download. Accessed August 3, 2016.
6. Streiner DL. Statistics commentary series: commentary #8 – effect sizes. *J Clin Psychopharmacol*. 2015;35:217–219.
7. Rosenthal R. The "file drawer problem" and tolerance for null results. *Psychol Bull*. 1979;86:638–641.
8. Cooper H, DeNeve K, Charlton K. Finding the missing science: the fate of studies submitted for review by human subjects committee. *Psychol Methods*. 1997;2:447–452.
9. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. *J R Stat Soc*. 1988;151:419–463.
10. Rothstein HR, Sutton AJ, Bornstein M, (Eds). *Publication Bias in Meta-Analysis – Prevention, Assessment and Adjustments*. New York: Wiley; 2005.
11. Joffe R, Sokolov S, Streiner D. Antidepressant treatment of depression: a metaanalysis. *Can J Psychiatry*. 1996;41:613–616.
12. Antman EM, Lau J, Kupelnick B, et al. Randomized controlled trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA*. 1992;268:240–248.
13. Canner PL. Aspirin in coronary heart disease. Comparison of six clinical trials. *Isr J Med Sci*. 1983;19:413–423.