Statistics Commentary Series

Commentary No. 19: Meta-Analysis, Part 2 – Some of the Issues

David L. Streiner, PhD, CPsych

In a previous article,[1] I described what meta-analysis (MA) is, briefly outlined how it's done, and pointed out some of its advantages. In that article, I also mentioned that MAs of the same topic can (and do) come to opposite conclusions because of different decisions they make along the way. For example, Greenberg et al.[2] and we[3] conducted MAs of the effectiveness of tricyclic antidepressants and our conclusions differed fundamentally. In this article, I will go back over some of the steps in an MA and discuss how decisions made at each of them may affect the final result.

1. *Which articles to include*. As mentioned in the first article, a hallmark of both systematic reviews and MAs is the thoroughness of the search and the inclusion of all articles that meet basic criteria (which may vary from MA to MA). However, the devil is in the details, and in this case, the detail is that small word "all." Not all studies finally appear in peer-reviewed journals. The results of some research are presented at meetings and never published; others as reports with limited circulation; and still others may have been theses or dissertations. This is referred to as the *gray literature* (or the "grey" literature in Canada and the U.K.), and the list may also include preprints, technical reports, discussion papers, drafts of articles in the bottom of file drawers, and the like. The question is whether these should be included. Those arguing against their inclusion believe that they have not been published for a very good reason – they are likely methodologically of poorer quality and that if they were sufficiently rigorous, they would have been published; in other words, they remain in the gray literature because they have or would have failed in the peer review process. On the other side of the debate, proponents for their inclusion point to the well-known publication bias mentioned in the previous article that results in mainly articles with positive results finding their way into press. This means that what has been published is a biased sample of all research on a topic, skewing the results in a positive direction. In fact, Kirsch et al. have written a number of papers showing that the effect sizes (ESs) for antidepressants reported in submissions to the Food and Drug Administration but not published were smaller than those in published articles.[4,5]

Complicating the picture even further is the unfortunate recent growth industry of "predatory journals." These are journals with very impressive titles (eg, *International Journal for Advanced Review and Research in Pharmacy*), often published overseas, that promise peer review and free access, but only after authors have paid a hefty fee. In fact, little if no peer review is done. One (of many) tests of the integrity of these journals consisted of a computer-generated paper filled with graphs, footnotes, and utter gibberish, and submitted from a fake institution (the Center for Research in Applied Phrenology; ie, CRAP) – it was accepted.[6] Fortunately, a research librarian, Jeffrey Beall, maintains an on-line list of predatory journals[7] and publishers.[8]

So, should articles from the gray literature or predatory journals be included? My personal position is that they should be, but with 2 provisions. First, the quality of the research should be evaluated using 1 of the existing scales, 25 of which were evaluated by Moher et al.[9] That allows you to use the score in a meta-regression to determine if the quality of the research affected the ESs. The second proviso is that the source of the article should be coded and entered into the data base (eg, 1 = peer-reviewed journal, 2 = predatory journal, 3 = gray literature). This can either be used as a dummy variable in a meta-regression, or the data analyzed both with and without the lesser-ranked articles. If these analyses show no effect of quality or the results are similar with and without those from the classes 2 and 3, then you can be fairly confident of the conclusions. However, if they do differ, it could be due to the fact that more poorly designed studies often have larger ESs than better ones[10,11] (in which case their inclusion would increase the ES), or that there is indeed some publication bias (and their inclusion would decrease the ES).

2. *Dealing with heterogeneity in ESs*. As mentioned in the previous article, the researcher should check for heterogeneity among the studies' ESs both statistically, with the *Q* statistic, and descriptively,

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.0000000000000667

using I^2. The problem is what to do about those differences if Q is significant or I^2 is greater than 50%. Not surprisingly, there are 2 different answers, with advocates on both sides of the issue. One side takes the position that, if heterogeneity is present, the MA is meaningless because it is trying to combine apples with oranges. That is, there must be substantive differences among the studies leading to different results, so that any 1 summary ES would be misleading. Proponents of this position would remove the study with the most discrepant ES, recalculate Q and I^2, and continue doing this until homogeneity is achieved. The other position, which is becoming the more popular practice, is to accept the heterogeneity and use sub-group analyses and meta-regression to try to determine the sources of the variability. Perhaps 1 reason for the increased acceptance of heterogeneity is the development of a statistical technique called *random-effects regression*, which will be discussed a bit later in this article. This is the approach that I would advocate.

3. *Dealing with other sources of heterogeneity.* There are other issues of differences among studies that must be considered earlier on in the MA process, in terms of setting criteria for which articles should be included in the first place. These have to do with differences in the definitions of the study participants, the treatments, and the outcome measures. For example, in an MA of antidepressants, should the criterion be depression, irrespective of sub-type or severity, or should the MA be restricted to only a specific group of depressed patients? Should the intervention consist only of selective serotonin reuptake inhibitors (SSRIs), or should it also include tricyclics and serotonin-norepinephrine reuptake inhibitors? Should the outcome consist of only clinician-rated scales or should it include self-rating scales too?

In part, the answer depends on the question. If the interest is solely in the effectiveness or efficacy of SSRIs, then of course the studies should be limited to only those drugs. Similarly, if the concern is their usefulness with elderly patients who have had a major depressive episode, then the selection criteria should reflect that. However, defining the question very narrowly precludes making possibly useful comparisons or statements about the generalizability of the findings, such as "this class of drugs is effective with these patients, but other classes are not," or "these interventions work with a wide range of patients." As with the criteria for the source of the articles and dealing with heterogeneous ESs, the wisest strategy is probably to be inclusive and see how these factors affect the ESs, through sub-group analyses and meta-regression.

4. *The analytic approach*. Earlier in this article, I mentioned that heterogeneity of ESs can be handled using a technique called *random-effects regression*. It's time now to bite the proverbial bullet and discuss how it differs from ordinary, fixed-effects regression. Most often when we run a multiple regression, say looking at the effect on quality of life (QOL) of a number of predictors, we assume that the strength of that effect is the same for everyone. For example, if we find that the b weight for 1 predictor, the number of previous depressive episodes, is 0.5, then the outcome (QOL) will increase by ½ a point for every episode of depression the person has had, and this relationship will hold true for every person in the population; that is, the effect is fixed for all people. However, it may be that the population consists of different subgroups (eg, men and women), and that the relationship is different in each. At the extreme, the relationship may be different for each person in the population; thus, rather than b being a fixed effect, it is *random*, varying from person to person.

Within the context of MAs, using a fixed-effects regression assumes that all of the studies share a common, true ES, and that the differences from 1 study to the next are reflections only of random variation. On the other hand, a random-effects model assumes that the true ESs vary from study to study, and the purpose of the MA is to estimate the average (not the true) ES. So which to use? We often see MAs or MA proposals state that they will use a fixed-effects regression unless the Q or I^2 statistic shows heterogeneity, in which case they will use a random-effects model, and it's easy to see why: the estimated ES is larger and the standard error smaller with a fixed-effects than with a random-effects model. Despite this, in my opinion, this is the wrong way to go. The type of analysis should be chosen *before* the data are analyzed, and should reflect the analyst's assumption about the nature of the ESs. I also believe that, in the vast majority of cases, the correct choice would be the random-effects model. Unless all studies use identical inclusion and exclusion criteria, have exactly the same intervention, follow patients for similar lengths of time, use the same outcome measure, and so forth, it is unlikely that they will all have the same true ES. Thus, the random-effects model is a better representation of the nature of the articles.

Summing up, conducting an MA is not like following the instructions from Ikea on how to assemble an article of furniture; it is closer to what a master woodworker does in building a custom-designed piece. There is a general pattern, but decisions can and must be made at each step, requiring thoughtfulness and at times ingenuity. The art is knowing what the options are and the ramifications of each.

AUTHOR DISCLOSURE INFORMATION

The author declares no conflicts of interest.

REFERENCES

1. Streiner DL. Statistics commentary series: commentary #18: Meta-analysis, Part 1 – What it is. *J Clin Psychopharmacol*. 2017;37:6–7.
2. Greenberg RP, Bornstein RF, Greenberg MD, et al. A meta-analysis of antidepressant outcome under "blinder" conditions. *J Consult Clin Psychol*. 1992;60:664–669.
3. Joffe R, Sokolov S, Streiner D. Antidepressant treatment of depression: a metaanalysis. *Can J Psychiatry*. 1996;41:613–616.
4. Kirsch I, Moore TJ, Scoboria A, et al. The emperor's new drugs: an analysis of antidepressant medication data submitted to the U.S. Food and Drug Administration. *Prevention & Treatment*. 2002; 5: Article 23://dx.doi.org/10.1037/1522-3736.5.1.523a.
5. Kirsch I, Deacon BJ, Huedo-Medina TB, et al. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLOS Med*. 2008;5:e45.
6. Basken P. Open-access publisher appears to have accepted fake paper from bogus center. *The Chronicle of Higher Education*. June 10, 2009. Available at: http://www.chronicle.com/article/Open-Access-Publisher-Appears/47717. Accessed January 6, 2017.
7. Available at: https://scholarlyoa.com/individual-journals/. Accessed January 6, 2017.
8. Available at: https://scholarlyoa.com/publishers/. Accessed January 6, 2017.
9. Moher D, Jadad AR, Nichol G, et al. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials*. 1995;16:62–73.
10. Sinclair J. Prevention and treatment of the respiratory distress syndrome. *Pediatr Clin N Am*. 1966;13:711–730.
11. Mansfield RS, Busse TV. Meta-analysis of research: A rejoinder to Glass. *Educ Res*. 1977;6:3.