EDITORIAL

An application of meta-analysis based on DerSimonian and Laird method

Brandon J. George, PhD,[a] and Inmaculada B. Aban, PhD[b]

[a] Office of Energetics, University of Alabama at Birmingham, Birmingham, AL
[b] Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL

See related article, pp. 680–689

In this issue, Elgendy et al discuss their findings from a meta-analysis of 22 studies that examined the clinical use of myocardial perfusion imaging (MPI) in situations not covered by the appropriate use criteria (AUC) put forth by the American College of Cardiology (ACC). In particular, the authors were interested in whether the inappropriate use of MPI resulted in different detection rates of cardiac ischemia or other abnormal findings compared to MPI used according to AUC.[1]

It is common in clinical research for an important research topic to have more than one study exploring that topic. There are many reasons for this, from replication and validation to assessing an effect or association in a different population. Meta-analysis allows researchers to compile the findings from different studies on a single topic in a structured, quantitative manner and use the joint knowledge of the field to make a more informed conclusion about a topic than from one study alone, or from multiple studies in a qualitative manner. As technology makes the aggregation of the research in a field more and more feasible, the scientific and funding communities are viewing meta-analysis as an efficient use of resources to get a definitive answer on a well-studied topic.

Conceptually, meta-analysis is similar to a typical study on individual patients. A standard study involves sampling subjects, where each subject has a particular

outcome (e.g., treatment effect) to be measured. A meta-analysis involves sampling studies on a topic, where each study has an aggregate outcome (e.g., a mean treatment effect) to be measured. In both cases, the outcomes from the sampling units are statistically compiled to produce an overall conclusion about that outcome in the population; for standard studies the population refers to the subject population, while for meta-analyses the population is all possible studies on that topic.

In standard studies, proper sampling methods are necessary so the sample is representative of the underlying population and selection bias is avoided. The same holds true for meta-analysis, where one wants the sample of studies to be representative of all possible studies on a subject. Unfortunately, the number of available studies on a topic is usually small and may be reduced further due to subject-specific exclusion criteria. Although Elgendy et al found hundreds of thousands of papers with a broad keyword search on MEDLINE, only 171 fit all of their relevant keywords; this 171 was further reduced down to 22 studies after manual review excluded papers that were not relevant, duplicates, or did not report usable data. It should be noted that one should take care to minimize the effects of publication bias in the literature search so that the sample of studies is truly representative of all studies done, not just those with favorable results; review of prospective registries (such as clinicaltrials.gov), conference proceedings, and technical reports may help identify studies that would have otherwise gone unnoticed.[2]

To further reinforce the parallels of standard studies and meta-analyses, it is common practice for meta-analyses to include a PRISMA flow diagram that details the search and inclusion of studies,[3] much like the CONSORT flow diagram details the flow of subjects into and through a clinical trial. A thorough discussion of how to select studies for a meta-analysis is outside the scope of this editorial, but those looking to perform a meta-analysis should consider the PRISMA guidelines

during the planning stage much like one considers the CONSORT guidelines when planning a clinical trial.[3] The methods put forth by the Cochrane Collaboration (www.cochrane.org) may also be of great help as well.

After a sample of studies has been selected, there are several statistical issues to address. The first is that unlike a clinical trial, in a meta-analysis each sampling unit is typically not considered to carry the same weight. It is reasonable to consider a study with a larger sample size or smaller variance of the outcome estimate to be more informative than a small or noisy study. To handle this, it is common to give varying weight the different studies in a meta-analysis. The most reasonable weights are the inverse of the variance of the estimate $N/\hat{\sigma}^2$, where N is the sample size of the study and $\hat{\sigma}^2$ is the variance of the outcome of interest reported by the study, which means that studies with larger sample size or smaller variance are given more weight. Although some researchers may consider a measure of the quality of a study to factor into its weight, such an approach is highly subjective and runs contrary to the goal of meta-analysis to provide an objective result.[4] Even if quality is not taken into account in the quantitative analysis, it is useful to examine and report the relevant characteristics of the studies' samples and methods, as in Table 2 in Elgendy et al.

Another important statistical issue is the handling of heterogeneity of outcome measures among the different studies in the meta-analysis, particularly when the subject-level outcomes are binary as in Elgendy et al where the two outcomes are normal/abnormal and ischemic/not ischemic MPI. Although we can assume that there is a universal value governing an association, defined as θ, the individual studies may have differences in materials, methods, or subjects (such as those listed in Table 2 of Engendy et al) that result in the true value for that study to be different by some amount, defined as δ_i for study i. For example, in Elgendy et al the estimate for the odds ratio, θ, of abnormal test results is 0.416; the estimated odds ratio from the first study (Winchester) was 0.215 so a simple estimate of δ_1 would be $0.215 - 0.416 = -0.201$. This heterogeneity can easily be accounted for in the analysis by including a random effect for study, whose statistical assumptions closely match our scientific assumptions. Although the actual deviations for each study will be non-zero, on average they will cancel out with respect to the true effect; statistically, we assume δ_i has a mean of zero. The variability in our estimate of θ, which is critical in constructing confidence intervals and hypothesis testing, is a function of both the heterogeneity among the studies a well as the sampling variance of each study; statistically we assume that δ_i has variance Δ^2 and is independent from the sampling error in each study. DerSimonian and Laird proposed an approximation method to estimate the value of Δ^2 that is easy enough to do in Microsoft Excel as well as a test for whether there is heterogeneity in the effect between studies.[5] The accessibility of the DerSimonian-Laird (DL) method and its inclusion in common meta-analysis software such as RevMan[6] has led to it being the most common method for using random effects in meta-analyses, and it is a fairly reliable approximation when the number of studies is large.[7]

Note that failure to account for the heterogeneity between studies will result in an underestimation of the variability of the overall effect θ[5] This means that leaving out the random effect leads to us assuming that we have a more precise measure of θ than we really do, resulting in an inflated rate of false positives. In the meta-analytic application in Elgandy et al, a false positive would be where one declares that the type of MPI test (appropriate vs inappropriate) indicates a statistically significant difference in the probability of an outcome (e.g., abnormal test or ischemia) when in fact none exists.

Despite the DL method's popularity, numerous papers have identified limitations in its use. It has been noted that the DL method can severely underestimate Δ^2 when the underlying proportion is near zero or one,[8] or if the number of studies is small (<20).[7,8] The method has also been found to produce an inflated false positive rate for the overall conclusion when there is a large variation in the sample sizes of the included studies; a tenfold difference in study size was found to produce results with very poor statistical properties using the DL method.[9] The majority of these issues are due to the original DL method being based on a simple approximation, which does behave well when the number of studies is large and the studies themselves are fairly uniform.[7] More sophisticated approximations have been proposed that have better statistical properties but can still be easily computed.[9,10] Research still consistently finds that more computationally intensive methods that utilize the full likelihood function have the best statistical properties,[5,7,8,10] although they require software such as SAS, the 'metafor' package in R, or Comprehensive Meta-Analysis 3.0 (which Elgendy et al used).

In their meta-analysis, Elgendy et al used a random effects term with estimation via the DerSimonian-Laird method. However, there are some properties of their data that may have made this a suboptimal method of analysis. Although 22 studies were included, they were split between the reported outcomes; analysis of abnormal test results only included eight studies while the one for ischemia included six studies. These are far

fewer studies than would be needed to have high reliability in the DL method, and thus the false positive rate for the overall conclusion may be inflated, meaning the significant finding of an association between AUC and abnormal test results or ischemia may be due to chance at a higher rate than 5%. This may be exacerbated by the heterogeneity in study size, with a large difference between the smallest (N = 206) and largest (N = 6351) studies. It is difficult to determine exactly how the results of the meta-analysis would change if re-analyzed with more advanced methods, particularly since the authors did not report the level of heterogeneity seen among the studies. To be fair, in this particular paper the overall effects are very strongly significant ($P < 0.001$ for abnormal test and ischemia) or not significant (for cardiologists vs not), so the results would probably not change qualitatively. However, if a meta-analysis reported a significant effect at $P = 0.045$ using the DL method in an inappropriate place, one should be wary of claims of statistical significance.

Meta-analysis provides a useful way to synthesize the results from an established field of research, but carries with it statistical challenges. In particular, the collection of studies must be done with care and the assumptions of the analytical methods must be assessed.

Disclosure

References

1. Elgendy IY, Mahmoud A, Shuster JJ, Doukky R, Winchester DE. Outcomes after inappropriate nuclear myocardial perfusion imaging: A meta-analysis. J Nucl Cardiol. 2015. doi:10.1007/s12350-015-0240-2.
2. Sutton A. Publication bias. In: Cooper H, Hedges LV, Valentine JC, editors. The handbook of research synthesis and meta-analysis. 2nd ed. New York: Russell Sage Foundation; 2009. p. 435-52.
3. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med. 2009;6:e1000097. doi:10.1371/journal.pmed.1000097.
4. Shadish WR, Haddock CK. Combining estimates of effect size. In: Cooper H, Hedges LV, Valentine JC, editors. The handbook of research synthesis and meta-analysis. 2nd ed. New York: Russell Sage Foundation; 2009. p. 257-77.
5. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177-88.
6. Review Manager (RevMan) [Computer program] Version 5.3. (2014) Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration
7. Jackson D, Bowden J, Baker R. How does the DerSimonian and Laird procedure for random effects meta-analysis compare to its more efficient but harder to compute counterparts? J Stat Plan Inference 2010;140:961-70.
8. Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. J Clin Epidemiol 2008;61:41-51.
9. IntHout J, Ionnidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014;14:25.
10. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: An update. Contemp Clin Trials 2007;28:105-14.