Meta-analysis as Evidence
Building a Better Pyramid

Jesse A. Berlin, ScD; Robert M. Golub, MD

In following the practice of evidence-based medicine, when faced with a question about prevention or treatment the clinician should seek out the best evidence that addresses the question. If quality of evidence is considered a pyramid, what category should be placed at the peak? One dogma argues that it is the best-conducted randomized clinical trial (RCT) comprising patients similar to those seen by the clinician, reasoning that a well-done RCT mimics pure experimental conditions better than any other study design, hence minimizing the likelihood of confounding. A counter-argument is that the best evidence is a systematic review with meta-analysis, because this approach can integrate all of the relevant evidence and provide a more reliable answer than a single study, however well conducted.

The notion that a synthesis that includes mathematically combining a complete body of evidence provides the highest level of evidence is attractive. However, as with most of evidence-based medicine, the principles are rational, consistent, and appealing, but in practice are fraught with practical challenges, ambiguities, and nuances. Moreover, a busy clinician faces tension between searching for and assessing the best-quality primary evidence vs accepting the efficiency of using easily obtained but potentially inferior information as a shortcut to an answer.

As a general principle, generating, summarizing, and understanding the best available evidence are essential for establishing the benefits and safety of interventions. Meta-analysis has become a valuable tool toward these ends. There has been a proliferation of guidelines by professional societies and others, aimed at ensuring that the best preventive interventions or treatment options are provided to the appropriate patients at the appropriate time; these guidelines often incorporate meta-analyses as a key evidence support for their recommendations.

However, limitations of meta-analysis as a study design preclude consistently placing this evidence at the top of the pyramid, and a number of issues need to be resolved before that can happen. These are the problems that researchers, guideline developers, journal editors, and critical readers of the literature struggle with, and understanding the limitations of metaanalytic evidence is crucial for each of these stakeholders. One useful way to view these challenges is to divide them into 2 categories: heterogeneity and methodological dilemmas.

Heterogeneity (variation in true effect sizes and in factors that might influence those effect sizes) is inherent in meta-analysis, not a problem to be solved. It includes clinical components (eg, diversity in patient populations or interventions) and statistical components (eg, random differences). There are statistical approaches to try to quantify some elements of heterogeneity, including the Q statistic (a measure of total within-study variance), the F statistic (the ratio of variability of results among studies to total observed variation), and ι² (a measure of between-studies variance). Heterogeneity can be investigated and sometimes managed, but not eliminated as an issue. In some instances, helpful insights can be gained when the heterogeneity of findings of component studies can be related to characteristics of those studies (eg, disease severity, outcome definition, duration of follow-up, duration of treatment, or dose of a drug). For example, treatments might show different effects in patients with severe disease than in those with mild to moderate disease. Studies might also give different answers because of flaws in the design or conduct, such as excessive loss to follow-up.

Modeling to understand heterogeneity can be helpful. However, relationships that would allow drawing conclusions about effects in different populations, or effects for an ideal study, do not always emerge from the analyses of heterogeneity. Study factors are often confounded with each other; for example, if high-dose studies were conducted in more severely ill populations and low-dose studies in less ill populations, it would be difficult if not impossible to separate the effects of dose from the effects of severity of illness. Most meta-analyses use the aggregate data as presented in the report of the primary study, but when individual patient-level data are not available important relationships can be missed and spurious relationships can be found.

Moreover, inherent heterogeneity presents a challenge in practical clinical interpretation, even when the heterogeneity is assessed and its effects understood. For a single RCT, interpretation of the average effect size may exaggerate the benefit that most patients are likely to achieve and often does not provide the information needed for a clinician to understand how to apply the findings to a particular patient. With a meta-analysis, in which 5, 10, or 40 RCTs are combined, the interpretation of the average effect size for application to an individual patient is likely to be even more obscured. Because of the innate uncertainty generated by combining RCT data from disparate sources (which can include differences in patients, interventions, and assessed outcomes), the ability to make causal inferences can be limited; JAMA considers meta-analysis to represent an observational design, with measures that should be interpreted as associations rather than causal effects.
In contrast to heterogeneity, methodological dilemmas are puzzles that are potentially solvable and that may in the future be eliminated as sources of uncertainty and error. There is often lack of clarity or consensus on how best to handle these areas. As a result, there may be conflicting recommendations for how to deal with them, and in the end articles may include a variety of approaches and analyses for transparency, yet leave the reader uncertain about the appropriate conclusion: this uncertainty is propagated into questions of how to interpret the meta-analyses.

These types of dilemmas are illustrated by the study by Dechartres and colleagues in this issue of JAMA.6 Using existing collections of publicly available meta-analyses, the authors asked the question of whether different approaches to including studies in a meta-analysis would lead to different estimates of effect size and different interpretations of the study findings. The key message of their analyses is that the approach a researcher uses will, in many cases, not only affect the magnitude of estimated effect associated with an intervention (such as treatment) but will in some cases change the findings from a statistically significant benefit to no association (such as treatment) but will in some cases change the findings from a statistically significant benefit to no association between the intervention and outcomes.

The authors compared the findings using 5 approaches: (1) including all the studies in the original meta-analysis; (2) using the result of the most precise trial (the one with the lowest statistical variability, i.e., the narrowest confidence interval); (3) conducting a meta-analysis of the 25% largest trials; (4) conducting a limit meta-analysis, which predicts the effect size for an infinitely large trial based on a statistical model of the available studies;7,8 and (5) comparing treatment effects between trials at high or unclear risk of bias vs trials at low risk of bias for each of the Cochrane risk of bias key domains of sequence generation, allocation concealment, blinding, and incomplete outcome data.9

The authors found that different approaches to providing the best evidence can lead to different results, making it unclear how to define “best.” Overall, treatment effects were larger in the meta-analysis of all trials than in the most precise trial. The difference in treatment outcomes between these strategies was substantial in 47 of 92 (51%) meta-analyses of subjective outcomes and in 28 of 71 (39%) meta-analyses of objective outcomes; however, in these comparisons with substantial differences, it was not always the meta-analysis of all trials that showed the larger treatment effect. There was a small difference between meta-analyses of all trials and meta-analyses of the largest trials, and a somewhat larger difference when comparing meta-analysis of all trials and limit meta-analysis. For 3 of the 4 domains for risk of bias, the treatment effects were larger for trials at high or unclear risk of bias compared with those at low risk (sequence generation, allocation concealment, and blinding) for both subjective and objective outcomes; however, there were no significant differences for the domain of incomplete outcome data.

The focus of many of these analyses was on study size or precision of effect size estimation. It was not a given that large trials would show smaller effects (although there are reasons to suspect that could be the case, including publication bias). It is also not always the case that larger trials are automatically better on any of the domains of risk of bias. But in the end, results were consistent for 3 domains of risk of bias (studies with low risk of bias generally showed smaller effect sizes) and precision-related aspects (larger studies generally showed smaller effects).

The problem of which studies should be included in meta-analytic pooling is just one of many methodological questions that potentially affect the quantitative results and the qualitative interpretation of meta-analyses but that have not been resolved in the research community. These include but are not limited to assessing publication bias; assessing risk of bias for specific domains in primary studies; appropriate statistical pooling techniques (a particularly important issue for rare events); and how to handle unpublished literature. Although the focus is often on meta-analysis of RCTs, for many questions it is necessary to instead use data from observational studies, for which considerations of within-study bias become even more important to understand, and at the same time more challenging to deal with, than in RCTs.

Ultimately, meta-analyses cannot always overcome the limitations of individual trials by pooling treatment effect estimates to generate a single best estimate. The original studies can still be biased, even if the meta-analytic methods are perfect. Meta-analysis can usually increase precision (reduce variability), but precision alone does not make up for bias in the component studies.

Moreover, it is not uncommon for there to be conflicting results between meta-analyses of the same topic.10 However, just as frequently, conflicting results also occur between large (and purportedly definitive) RCTs on the same topic.11 There can be many reasons for these discrepancies (e.g., individual studies might include different patient populations or use different outcome definitions; different meta-analyses might include different sets of studies), but the lesson is that it is not always immediately clear which result is correct, or even what standard to use to define correct.

Meta-analysis may also be particularly prone to a third type of problem: conduct by researchers without methodological expertise. It is relatively easy to create a mediocre or poor-quality meta-analysis: an expert literature search and data extraction require few resources and little training, and contemporary statistical packages (including those that are free) make it possible to complete an entire study quickly. This makes meta-analysis attractive to researchers with limited training or mentoring and has led to a plethora of such publications; the annual number of PubMed publications indexed under “meta-analysis” as publication type increased from 1289 in 2003 to 7053 in 2013. Many of these are of dubious quality and address questions of limited importance. This is likely to become a more prominent issue as data-sharing becomes more common and individual participant-level data become increasingly available. As with other research methods, properly conducted meta-analyses require clinical and technical expertise involving cross-disciplinary teams,1 including skill in literature search, data extraction, and statistical pooling.

What can be done to improve the quality of the science and the value of meta-analytic evidence? First, the importance of heterogeneity needs to be recognized and adequately ex-
explored in all meta-analyses for which this is a factor. Reporting statistics such as F or r² is not sufficient. Authors should conduct individual patient-level meta-analyses whenever possible, aggregate data-level meta-regression when appropriate, and subgroup analyses for all relevant sources of heterogeneity. Interpretations need to be nuanced and must incorporate the insights gained from such explorations.

Second, researchers should recognize the complexity of conducting a high-quality meta-analysis. The research team should include members with expertise in each step of design, execution, and interpretation of a systematic review and meta-analysis. In addition, there are several guidelines for conducting meta-analyses. Third, as with RCTs, there should be a protocol for the systematic review and meta-analysis, in which all the analyses are specified in advance of knowing the results of those analyses. These protocols should be prospectively registered at a site such as PROSPERO to ensure fidelity in study execution as well as provide a database of meta-analyses to help future researchers. Deviations from the protocol, and their necessity, must be explained so that readers can assess their implications. Decisions (which need to be documented) are made at every step of every meta-analysis; in addition to documenting these decisions, it is incumbent on meta-analysts to perform sensitivity analyses in which the robustness of results is tested against alternate approaches and assumptions.

Fourth, and perhaps most important, the published form of the article must include all of the information that readers need to completely understand how the study was conducted and to have all of the data available to independently assess the validity of the analyses and to reach their own interpretations of the findings. Formal guidelines for reporting these studies (such as PRISMA) are available; these emphasize the need for transparency and completeness in reporting methods and results of systematic reviews and meta-analyses. Although these are guidelines for how to report study findings, if researchers use them to help design their study it is more likely that an important element will not be missed. If researchers use the guidelines to write their manuscripts, it makes the peer review process more efficient and more informed. But ultimately, it is the responsibility of the journals that publish meta-analyses to make certain that their articles adhere to these guidelines as best they can.

Findings such as those in the study by Dechartres et al reinforce concerns that journals and readers have about meta-analysis as a study design. Those findings deserve consideration not only in the planning of the studies but in the journal peer review and evaluation. They also reinforce the need for circumspection in study interpretation.

Meta-analysis has the potential to be the best source of evidence to inform decision making. The underlying methods have become much more sophisticated in the last few decades, but achieving this potential will require continued advances in the underlying science, parallel to the advances that have occurred with other biomedical research design and statistics. Until that occurs, an informed reader must approach these studies, as with all other literature, as imperfect information that requires critical appraisal and assessment of applicability of the findings to individual patients. This is not easy, and it requires skill and intelligence. Whatever clinical evidence looks like, and wherever it is placed on a pyramid, there are no shortcuts to truth.

ARTICLE INFORMATION
Author Affiliations: Johnson & Johnson, Titusville, New Jersey (Berlin); Deputy Editor, JAMA (Golub).
Corresponding Author: Robert M. Golub, MD, JAMA, 330 N Wabash Ave, Ste 39300, Chicago, IL 60611-5885 (robert.golub@jamanetwork.org).
Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.
Disclaimer: Dr Berlin is Vice President and Global Head of Epidemiology for Johnson & Johnson. The views expressed herein do not necessarily represent the views or practices of Johnson & Johnson or any other party.
Correction: This article was corrected online on October 6, 2014, to correct author names in the reference list.

REFERENCES