Editor’s Commentary: Additional Associate Editors, New Submission Category, and Meta-methodology

As I write this column, I have recently returned from attending the 23rd Annual Euro-Meeting in Geneva. Our sister publication, the Global Forum, will report on this successful event in more detail, but I would like to take the opportunity to acknowledge an important scientific component of the meeting, the awarding of the prizes in the student and professional poster categories. Our colleagues in the European DIA offices in Basel organized two wonderful ceremonies, held at the DIA booth in the Exhibition Hall. The winner of the student award was Hans Ebbers. His abstract, titled “Determinants of Safety Related Regulatory Actions of Biopharmaceuticals,” appeared in the March issue of the journal, along with other entries in this category. The winners of the professional award were Pasi Korhonen, Jari Haukka, Mark Taylor, Peter Haddad, Maxine Patel, and Jari Tiihonen. Their abstract was titled “Rehospitalization Risk and Discontinuation of Initial Antipsychotic Treatment After First Hospital Episode of Schizophrenia,” and the quality of their work is reflected in its acceptance for publication in the American Journal of Psychiatry (1).

Also, while mentioning meetings, this issue of the journal contains reports of two meetings, one held in London in January 2010 and the other in Washington, DC, in January 2011. We are keen to be able to provide reports of more such meetings. Therefore, if you would like to prepare a report of a meeting or conference that you organized or attended, please contact our editorial office to discuss this possibility.

APPOINTMENT OF ADDITIONAL ASSOCIATE EDITORS

In the March issue of this volume I had the privilege of introducing Peter Pitts and Ken Getz as associate editors, and had equal pleasure in introducing Todd Durham in the previous issue. I am now very pleased to introduce two more associate editors to round out this new group. Jennifer Christian, PharmD, MPH, PhD, is a manager in the Cardiovascular and Metabolic Epidemiology Group at GlaxoSmithKline. Among her publishing and editorial experiences are leading the preparation of a Cardiac Safety Research Consortium White Paper titled “Cardiac Imaging for the Evaluation of Drug-induced Cardiotoxicity,” which will be published in a series of such articles in the American Heart Journal, and being the editor of Scribe, the newsletter of the International Society for Pharmacoepidemiology. Edward Tabor, MD, vice president of regulatory affairs at Quintiles, is our new book review editor. An acknowledged expert in drug-induced liver injury and a prolific author himself (he has over 300 publications), Ed will write many of the reviews himself, occasionally calling upon colleagues to do so too. Ed and I invite all readers who would like their own book to be reviewed to invite their publisher to send a copy to Ed via the DIA’s American headquarters in Horsham, PA. Additionally, if you come across a book that you feel is particularly worthy of consideration for review, please let me know and I will contact the book’s publishers on your behalf.

It was noted in my first Editor’s Commentary
that associate editors would be chosen to contribute a range of academic, government, industry, and not-for-profit perspectives within medical product development and interventions, and the clinical, statistical, public health, and regulatory sciences that facilitate such health-related endeavors. These individuals fulfill this mission particularly well, and I am pleased that they are already acting as ambassadors for the journal by raising its visibility in their respective professional circles.

NEW SUBMISSION CATEGORY

In addition to the several new categories of contributions to the journal that have appeared already, we would like to announce the Rapid Communications category. The goal of this type of submission, which will be peer-reviewed, is to facilitate the opportunity to publish empirical research findings in a rapid manner: both review of submissions and publication of accepted manuscripts will be expedited. Guidelines for content are a limit of around 1,750 words, two to three figures and/or tables, and around five key references. Since articles appearing in this category will be peer-reviewed publications, they should be regarded as the single report of the original data presented: more detailed and lengthy versions will not be considered. However, this vehicle may be ideal in certain circumstances, such as concise reports of preliminary data from pilot studies, and the sole published (and citable) report of empirical data validating a new measurement methodology against an accepted gold standard. We look forward to receiving these manuscripts.

META-METHODOLOGY: META-ANALYSIS REFRAMED

Meta-analyses have attained increasing prominence in the evidence-based medicine literature. Unfortunately, however, following the publication of their article in a journal, some meta-analysts disseminate their findings in the mass media with a bravado that markedly departs from calm, scientific, and clinical discourse, and seemingly with the expectation that the nation’s physicians will change their practice of medicine immediately. As Turner et al. (2) noted, “In the era of sensationalist, sound-bite coverage, clinical science sadly falls very low on the list of points to be covered in the allotted 30 seconds of television coverage,” a point not unknown to meta-analysts who willingly participate in this circus. Fortunately, many more authors take a more circumspect approach, providing physicians and their patients with well-reasoned and appropriately presented benefits and potential harms of a given intervention in the spirit that treatment decisions be made, as they should be, by physicians and their patients on a case-by-case basis. The intervention is made when physician-patient agreement is reached that the benefit-risk balance is favorable.

The term meta-analysis, while at first appearing appropriately descriptive, does not adequately capture the need for methodological rigor (5) in the full array of required actions. Certainly, an analysis is conducted, but determining the data set, choosing the appropriate analytical strategy, interpreting the numerical results of the analysis in the context of the research question being asked, and consistently presenting the results with scientific and clinical decorum are also critical.

In the case of an analysis of results from a collection of randomized clinical trials, the methodological rigor required by each individual trial is also required for all aspects of facilitating, conducting, and reporting the new analysis. Just as each trial had a study protocol and a statistical analysis plan (or a statistical section in the protocol) written before its commencement, a similar approach is required here too. It is therefore suggested that the term meta-methodology more meaningfully reflects the research technology employed. As in experimental methodology, the optimum-quality answer to the research question of interest is dependent upon the employment of optimum-quality study methodology and appropriate statistical analysis that is dependent upon the nature of the data themselves.
Meta-methodology facilitates a quantitative evaluation of the evidence provided by two or more individual trials that have addressed the same research question. It commonly involves the statistical combination of summary statistics from various trials (study-level data), but it also refers to analyses performed on the combination of subject-level data. Reasons for the employment of meta-methodology include the following:

- Providing a more precise estimate of the overall treatment effect of interest (the treatment effect can be in the efficacy or the safety realm)
- Evaluating whether overall (positive) results are also seen in prespecified subgroups of participants
- Evaluating an additional efficacy or safety effect that requires more power than any of the individual trials incorporated can provide
- Evaluating an effect in a subgroup of participants, or a rare adverse event in all participants
- Assessing the possibility of a systematic effect among apparently conflicting study results

The conceptual basis of meta-methodology is straightforward: more data provide a better opportunity to get an optimum-quality answer to a research question. However, as alluded to previously, appropriate selection of studies to be included, implementation of the appropriate statistical techniques, and the appropriate interpretation and communication of the results obtained are not so straightforward.

The technique has both strengths and weaknesses, and both advocates and detractors. Turner and Durham (4) commented as follows:

If all of the components involved in conducting a meta-analysis are performed appropriately, and the extent to which the results are helpful is not overstated (that is, any limitations are appropriately acknowledged and shared whenever and wherever communicating the results), the results can be informative and instructive. Unfortunately, however, it is easier than one might suspect to conduct a meta-analysis inappropriately and then to overstate the results in a variety of circumstances.

Even though meta-methodology does not require a new trial to be conducted, it is still a research method in its own right. Therefore, as noted by Kay (5), “to ensure that a meta-analysis is scientifically valid, it is necessary to plan and conduct the analysis in an appropriate way. It is not sufficient to retrospectively go to a bunch of studies that you like the look of and stick them together!”

The purpose for a meta-methodological study must be stated at the outset, and the treatment effect of interest determined and identified. Subsequently, the basic steps are as follows:

- Establishing rules for which of the studies about to be identified will be incorporated
- Identification of all pertinent studies
- Data abstraction and acquisition
- Data analysis
- Evaluating heterogeneity
- Evaluating robustness
- Dissemination of results and conclusions

One straightforward approach is to include every study identified and obtained. A counterargument is that, almost certainly, some studies will be “better” than others, and that “less good” studies should perhaps not be included if optimum-quality information is to be provided to the clinicians who will read publications of the results and therefore may base treatment decisions on these results. In the latter case, a priori inclusion and exclusion criteria that operationally define better and less good must be stated in advance of searching for studies, thereby determining which of the studies about to be identified will be included in the analysis.

Identification of all studies that may potentially be included has become much easier with the advent of computer search engines and web-based tools, but it can still be a challenging task, particularly when trying to locate and obtain unpublished data (publication bias is a powerful potential confounding factor in this context).

Data extraction and acquisition also has its challenges. Many published reports of individual trials present summary statistics as opposed
to presenting the underlying data, that is, the subject-level data. While Piantadosi (6) noted that “more than the published information is usually necessary to perform rigorous overviews,” this requires the meta-analysts to obtain and analyze subject-level data (not simply aggregate data across treatment groups) from each study. Unless a sponsor is conducting an analysis of data from several of its own studies, as can now be the case during the development of new antidiabetic drugs for type 2 diabetes mellitus (7,8), it can be difficult to gain access to subject-level data for every study of interest. For this reason, meta-analyses based on summary measures of the treatment effect gathered from published studies are currently more commonly reported. Accordingly, discussions in this column pertain to such analyses.

**DATA ANALYSIS: FIXED-EFFECT AND RANDOM-EFFECT MODELS**

The goal of meta-methodology is to estimate the overall treatment effect of interest across all of the studies included, that is, to obtain a single estimate of the treatment effect and the variance associated with it. Two items of data are obtained from each study incorporated:

1. A measure of the treatment effect in that specific study
2. An estimate of the variance associated with that specific treatment effect

However, before conducting the analysis, it must be decided whether to employ a fixed-effect analysis model or a random-effect analysis model. The fundamental difference between the models pertains to the degree of influence each individual treatment effect is allowed to exert mathematically on the overall treatment effect.

Each study included in the analysis contributes the same piece of information to the analysis, that is, the treatment effect found in that study. Therefore, if the analysis incorporates 100 studies, 100 treatment effects are included. However, each item of information does not necessarily carry the same weight when determining the result of the analysis: some items are accorded more influence than others. This influence is expressed in terms of the weight ascribed to each study and hence to each individual treatment effect. Studies whose treatment effects are weighted more heavily will exert a greater influence on the final result of the analysis than those whose treatment effects are weighted less heavily. The weighting accorded to each study-specific treatment effect is determined computationally according to the rules of the analysis model adopted.

In the fixed-effect model, the weight assigned to each study is the inverse of the variance associated with the study’s treatment effect. Therefore, the more precise the treatment effect estimate (ie, the smaller its variance), the greater the study’s influence on the overall treatment effect. Larger studies tend to yield more precise estimates, so this weighting system suggests that the overall estimate will be more consistent with those reported in the larger studies incorporated. In the random-effect model, the calculation of the weight assigned to each study involves an additional component. The weight assigned to each study is a function of the inverse of the variance of the treatment effect estimate (the sole item used in the fixed-effects meta-analysis model) and an estimate of the between-study variance in underlying parameters, since, unlike the fixed-effect model, the random-effect model acknowledges that these studies can vary.

This is an intuitively sensible feature of the random-effect model, since in most circumstances it would be very surprising if the studies incorporated did not vary from each other. This variation arises from the fact that each study included in the analysis will likely have many unique features, such as the nature of the study population and the number of subjects in each treatment group within the study, the length of treatment periods, the investigational sites at which the study was conducted, the quality of measurements made, and perhaps even in the precise way the treatment effect of interest was defined and therefore calculated.

While the methods of estimating this additional component need not be discussed here,
it is important to recognize the consequences. The more similar the individual studies included in the analysis, the less the impact of this component. That is, the more similar the studies, the closer the results obtained from the two analysis models: if there were no difference between studies, the results generated from the two analysis models would be identical. Phrasing this the other way around is perhaps more helpful: the greater the degree of difference between the studies incorporated in the analysis, the more important the choice of analysis model becomes.

The salient advantage of the random-effect model is that it allows between-study variability to influence the overall treatment effect estimate and, more particularly, its precision. This precision is captured by the estimate of its variation, which is represented by confidence intervals placed around the point estimate of the overall treatment effect generated by the analysis method chosen. The random-effect model tends to generate wider confidence intervals, indicating less precision in the point estimate. The consequence of this observation is as follows. If a fixed-effect model (which assumes that the studies included in the analysis do not differ) is employed when there is indeed a considerable difference between the studies included, the confidence placed in the result of the analysis will be greater than it should be. Had a random-effect model (which takes into account differences between the studies) been employed, the confidence interval placed around the point estimate would have been wider, indicating less confidence in the precision of the result obtained.

The relevance of the width of a confidence interval placed around a point estimate is that, for a given point estimate, it is possible that a more narrow confidence interval placed around it would result in a statistically significant finding from the analysis whereas a wider confidence interval would not. While the clinical significance of the result from such an analysis focuses much more on the confidence intervals placed around the point estimate than on the point estimate itself, achievement of statistical significance alone often drives the media’s interest in the result.

**EVALUATING HETEROGENEITY**

The statistical theory underpinning such analysis assumes that the study-specific estimates of the treatment effect are (relatively) homogenous. Homogeneity is present when the study-specific estimates are similar in magnitude and direction to the estimate of the treatment effect resulting from the combined analysis. Heterogeneity can arise from differences between studies, such as the possibilities already noted. Since the objective is to calculate a well-justified combined estimate of the treatment effect of interest, a formal evaluation of homogeneity following a visual graphic inspection of the combined effect against each individual effect is a recommended strategy.

This formal evaluation involves a statistical test, such as the Cochran $Q$ test. Homogeneity (also expressed as lack of heterogeneity) is indicated by a statistically nonsignificant result. While general acceptance of the $\alpha = 0.05$ criterion provides a line in the sand that is useful in certain circumstances, blind adherence to characterizing a result as either statistically significant or not statistically significant using the $\alpha = 0.05$ level is not necessarily a clinically meaningful strategy. In this context, a statistically nonsignificant Cochran test can be (mis)interpreted to state that there is no heterogeneity present. That is, a fallacious argument can be made that the lack of statistically significant evidence of heterogeneity represents an all-or-none statement of its complete absence.

**EVALUATING ROBUSTNESS**

Having calculated the result of the analysis, it can be informative to assess its robustness. In any combined analysis, some of the studies included will be larger than others, and sometimes a small percentage of included studies can be considerably larger than the majority of others. As already noted, the nature of the calculations performed here means that the larger trials tend to influence the result more, since they tend to have greater precision.
It can therefore be helpful to assess the robustness of the overall conclusion by performing the analysis without the data from the largest study or studies to see if the results remain qualitatively the same. If they do, then the result of the primary overall analysis is deemed robust. If they do not, confidence in the overall result can be undermined. Moreover, if the results are considerably different, it simply may not be appropriate to present the combined result alone and make statements based on it.

While the word qualitative may initially seem surprising here, qualitative assessments can complement primary quantitative assessments in various instances. Another example of qualitatively different results obtained with various analytical approaches leading to a lesser degree of confidence in the interpretation of results from a study is seen in the analysis of efficacy data presented in preapproval clinical trials. Data from two analysis populations, the intent-to-treat (ITT) and the per-protocol populations, are typically used in two separate sets of efficacy analysis. Regulatory agencies are generally much more encouraged if the two sets of results are qualitatively similar (the probability of them being quantitatively identical is vanishingly small). If they are not, questions may be raised as to why they are not. Additionally, if the per-protocol population is a lot smaller than the ITT population (it will almost certainly be somewhat smaller), regulatory reviewers will wonder why. Were there a lot of major protocol violations? Were a lot of subjects removed for the same protocol violation? Were many of the subjects with protocol violations enrolled at the same investigational site? Are there any systematic problems in the conduct of the trial? All of these issues can reduce overall confidence in the trial’s findings (9).

HYPOTHESIS GENERATION AND HYPOTHESIS TESTING

The discipline of statistics, like the practice of medicine, can be described as a science and an art. Performing numerical calculations is the easy part: knowing how best to design a study to answer a research question, collecting the appropriate data, and then analyzing them using the appropriate statistical procedures is the artful and skillful part. Among many skillful actions is strict adherence to the statistical convention dictating that a set of data generating a research hypothesis cannot be used to test that hypothesis. As one example of where this convention is important, consider subgroup analysis conducted using data from a large clinical trial. The response of well-defined subgroups of subjects to a therapeutic drug is a topic of legitimate medical interest. It may well be biologically plausible that, on average, members of certain well-defined subgroups would respond differently than members of other well-defined subgroups or from the study population as a whole, meaning that efficacy and/or safety concerns may be quite different for them. Therefore, it is clinically legitimate to address this question.

Investigation of such differences may be stated a priori as a secondary endpoint analysis in a trial’s protocol or statistical analysis plan. It is also possible that a biologically plausible subgroup difference that had not been anticipated may be identified in post hoc analysis, thereby generating a research hypothesis for subsequent testing. The key point is that such findings are best thought of as hypothesis generating (creating a question) and not hypothesis testing (answering a question) (3). To answer a question arising in this matter, it is appropriate to conduct a subsequent trial where the item of interest is the primary endpoint analysis, with the study powered appropriately to answer the question (test the hypothesis). With regard to meta-methodology, if a research question is generated by inspection of a given data set from a given trial, those data should not be included in a combined analysis data set used to test the hypothesis generated: to do so is to operate in the methodological twilight zone.

Going a step further, a result from a combined analysis is best thought of as hypothesis generating, that is, creating a question, than hypothesis testing, that is, answering a question (3).
DISSEMINATION OF RESULTS AND CONCLUSIONS

The results of a completed analysis would typically include the following:

• The estimated treatment effect (a point estimate) for each individual study included in the analysis and a confidence interval (often the 95% interval) about each study’s estimate
• The overall estimated treatment effect and its confidence interval

This information can be displayed in tabular form, or in a graphic form called a confidence interval plot.

As is true across all research methodology, if the correct study design has been employed and rigorous methodology has permitted the acquisition of optimum-quality data, the computational analysis itself is typically not difficult. The results are therefore not difficult to calculate. As implied earlier, the maturity and intellectual honesty of the researchers lies in the interpretation of the results and the appropriate degree of restraint needed to disseminate their conclusions in a responsible manner. Given all of these considerations, meta-methodology must be undertaken carefully, diligently, and responsibly.

In summary, while the term meta-analysis is widely used and accepted, this column has presented a case for the use of the term meta-methodology since the term meta-analysis does not adequately capture the need for methodological rigor in the full array of actions required for its meaningful execution. Perhaps both terms can be meaningfully used, with meta-analysis referring to the combined analysis conducted once necessary preceding methodological aspects of conducting this form of investigation have been completed. The term meta-methodology may prove to be a useful reminder that there is much more to this form of investigation than some mathematically correct calculations.

REFERENCES
