META-ANALYSES have gained popularity in medicine recently as a tool for summarizing evidence to help make medical decisions. Until now, large, single randomized, controlled trials have traditionally been considered the "gold standard" for the evaluation of medical technology and interventions, but on most medical questions it has been possible to perform only relatively small trials. Yet, do the results of meta-analyses of small trials agree with the results of large trials? Can meta-analyses validly substitute large trials? These questions are important to answer, to evaluate the utility of meta-analyses in clinical practice, and to understand the frequency and sources of discrepancies between meta-analyses and large trials. Such discrepancies have created controversy over the relative merits of both research methods and uncertainty about what evidence clinicians can rely on to make decisions.

Three different groups of investigators recently addressed these issues systematically using different protocols for selecting studies and comparing the 2 methods. Seemingly discordant conclusions led to further controversy. Villar et al reported moderate agreement, beyond chance, between meta-analyses and large trials. Cappelleri et al concluded that the 2 methods usually agree, and plausible reasons may be found for most discrepancies. Finally, LeLorier et al concluded that a meta-analysis fails to predict accurately the results of a subsequent large trial 35% of the time. In this communication, we evaluate systematically these investigations, so as to better understand their differences beyond the potential of emphasis on preconceived opinions, and to identify common themes in the comparison of meta-analyses with large trials that would enhance the medical community's appreciation of both research methods for informing clinical practice.

DESIGN OF PROTOCOLS TO COMPARE META-ANALYSES WITH LARGE STUDIES

Selection of Studies

The selection of meta-analyses and large trials varied substantially in the 3 protocols (Table 1). The Villar and Cappelleri protocols selected published meta-analyses where at least 1 large trial was included, and compared the large trial(s) against the remaining smaller trials. The LeLorier protocol identified large trials and then searched MEDLINE to identify published relevant meta-analyses. The former approach, where large trials had already been combined with smaller trials in meta-analyses, might select for more limited heterogeneity since formal pooling in meta-analyses may be avoided when large heterogeneity is discerned. The advantage of selecting previously performed meta-analyses is that the comparability of large and small studies has already been judged by meta-analysts, while the LeLorier protocol had to subjectively judge that meta-analyses and subsequent trials had addressed comparable questions.

Villar used the Pregnancy and Childbirth Module of the Cochrane database. This is the most systematically reviewed discipline in medicine to date. Thus, there was probably no selection bias against medical questions where the accumulated evidence had not been subjected to meta-analysis. However, generalizing to other disciplines where meta-analyses are performed more sporadically is not ensured. Cappelleri used an expanded version of the same database and added MEDLINE-indexed meta-analyses. This increased generalizability, but may have selected more homogeneous questions where meta-analyses were deemed appropriate. Finally, the clinical domains addressed by LeLorier comprised a set with binary outcomes (a subset of the Cappell-
Table 1.—Design of Protocols Comparing Meta-analyses With Large Trials

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Villar et al(^{a})</th>
<th>Cappelleri et al(^{b})</th>
<th>Cappelleri et al(^{c})</th>
<th>LeLorier et al(^{d})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcomes of trials</td>
<td>Binary</td>
<td>Binary</td>
<td>Binary</td>
<td>Binary or continuous</td>
</tr>
<tr>
<td>Large trials, (n)</td>
<td>30</td>
<td>178</td>
<td>175</td>
<td>12</td>
</tr>
<tr>
<td>Comparisons of large trials with meta-analyses, (n)</td>
<td>30</td>
<td>79</td>
<td>61</td>
<td>40</td>
</tr>
<tr>
<td>End points addressed</td>
<td>Primary</td>
<td>Primary</td>
<td>Primary</td>
<td>Primary and secondary</td>
</tr>
<tr>
<td>Definition of large trial</td>
<td>(\geq 1000) Patients</td>
<td>(\geq 1000) Patients</td>
<td>(\geq 80%) Power</td>
<td>(&gt; 1000) Patients</td>
</tr>
<tr>
<td>Large trials per comparison</td>
<td>(1^{*})</td>
<td>All trials qualifying as large</td>
<td>All trials qualifying as large</td>
<td>(1^{†})</td>
</tr>
<tr>
<td>Large trials already included in meta-analyses</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Large trials last or next-to-last published among studies</td>
<td>21/30</td>
<td>. . .‡</td>
<td>. . .‡</td>
<td>12/12</td>
</tr>
<tr>
<td>Large trial(s) had fewer patients than all small studies combined</td>
<td>19/30</td>
<td>27/79§</td>
<td>16/61</td>
<td>6/12</td>
</tr>
</tbody>
</table>

\(^{a}\)The largest trial.

\(^{b}\)The latest trial but not necessarily the largest.

\(^{c}\)Data not applicable because there could be several large trials considered in the Cappelleri protocol.

\(^{d}\)In 13 comparisons (not included in the 27), all the large trials combined had more patients than the smaller trials combined, but the largest trial had fewer patients than all the smaller trials combined.

Figure 1.—Randomized controlled evidence seen in the context of trial findings and trial sample size and the effect of these factors on the selection of trials for meta-analyses (used for identifying large trials in the Villar protocol\(^{a}\) and the Cappelleri protocol\(^{b}\) and for publication in influential journals (used for identifying large trials in the LeLorier protocol\(^{d}\)).

Two approaches for defining a large trial were used. All 3 protocols used a rule based on sample size, with 1000 patients or more qualifying as a large trial. Cappelleri also used a second approach based on statistical power where a trial is large if, given the proportion of the event of interest in the control group of smaller trials, it has 80% power to detect the treatment effect suggested by smaller trials (\(\alpha = .05\)). The Villar and the LeLorier protocols allowed for only 1 large trial. Cappelleri allowed for multiple large trials if they satisfied the corresponding definition.

Definition of Large Trial

Two approaches for defining a large trial were used. All 3 protocols used a rule based on sample size, with 1000 patients or more qualifying as a large trial. Cappelleri also used a second approach based on statistical power where a trial is large if, given the proportion of the event of interest in the control group of smaller trials, it has 80% power to detect the treatment effect suggested by smaller trials (\(\alpha = .05\)). The Villar and the LeLorier protocols allowed for only 1 large trial. Cappelleri allowed for multiple large trials if they satisfied the corresponding definition.

The sample size rule is convenient and simple, but does not capture important trial elements such as the event rates, follow-up, and complexity of design and execution. A trial of 1000 patients with 1:1 randomization has approximately 80% power (\(\alpha = .05\)) to detect a reduction in the frequency of the outcome of interest from 10%
to 5%. Nevertheless, a 1,000-patient study may still be grossly underpowered if the event studied is uncommon or the treatment effect is small, but still clinically meaningful to detect. For example, cardiology trials often seek reductions of the 30-day mortality rate after myocardial infarction from 7% to 6%. Moreover, large trials most often are underpowered for secondary end points and subgroup analyses. Power in large trials may further be eroded from missing data, losses to follow-up, and null bias, which is the serendipitous use of experimental treatment in the control group. Null bias is more common in simple design large trials than in smaller trials with rigid definitions of the experimental conditions. Large treatment effects observed in small, selective trials may shrink when the same therapies are used in community settings.

Even more sophisticated power-based approaches do not account for problem-specific peculiarities in the design and execution of a study. Studies on the same problem may require largely different levels of intensity of effort for their conduct. The power approach also tends to select meta-analyses where small studies have large treatment effects and high event rates, since under such circumstances, a trial would qualify as large with fewer patients. The effect of this selection shift in the frequency of observed disagreements is uncertain. Finally, neither definition of a large trial captures the quality of the trial’s design and execution.

The caveats mentioned above highlight the difficulty of separating large and small trials. The situation is even more problematic because a meta-analysis often includes several large trials and the gold standard large trial(s) have fewer patients than the meta-analysis of small trials (Table 1).

### Measures of Agreement Between Meta-analyses and Large Trials

Agreement between meta-analyses and large trials may be assessed either on the level of nominal statistical significance or on the magnitude of the treatment effect.

One approach assumes the most important aspect of the results of a trial or a meta-analysis is whether the P value is less than or greater than .05. Simply put, the treatment either works or it does not work. The second approach focuses on the magnitude of the treatment effect. For example, an odds ratio (OR) of 0.78 (confidence interval [CI], 0.42-1.45; P = .43) and an OR of 0.78 (CI, 0.64-0.95; P = .01) would be considered discordant with the former approach; with the latter approach, results differing in precision, but agreeing in point estimates, are concordant.

### Agreement in Level of Statistical Significance

Villar used 3 categories: agreement, when large trials and meta-analyses agreed in the direction of the treatment effect and both were significant or both were nonsignificant; partial agreement, when the 2 methods agreed in the direction of the treatment effect, but not in level of significance; and disagreement, when the 2 methods showed treatment effects in opposite directions. LeLorier merged partial disagreements with disagreements.

The classification scheme affects the estimated overall concordance. An explicit classification could use 3 classes: significant in favor of the experimental arm, nonsignificant, and significant against the experimental arm. The middle category could also be split in two, depending on whether the direction is in favor of or against the experimental arm. The level of agreement, above and beyond chance, between variables in a classification table may be evaluated with a kappa coefficient, which may either consider partial agreement (weighted kappa) or count only perfect agreements (unweighted kappa). Table 2 shows the estimated predictive performance and kappa coefficients for the different protocols when different classifications are used. Estimates of kappa range widely from 0.22 to 0.72 and 95% CIs extend from no agreement (−0.06) to excellent agreement (0.93). Unweighted kappa coefficients underestimate weighted ones where finer levels of disagreement are allowed. This is not surprising since, even when meta-analyses disagreed with large trials in direction of effect, in no case was statistical significance claimed in opposite directions. Notably, the LeLorier protocol shows lower kappa values compared with the other 2 protocols, probably because of differences in study and endpoint selection.

### Correlation of Treatment Effects

Standardized correlation coefficients between the magnitude of the treatment effect (how much the treatment works) in large trials and in meta-analyses have the advantage of avoiding the subjective classification of significance, agreement, and disagreement. The disadvantage is that they do not provide a clinical impression of the extent of agreement, but are based on an overall estimate of the treatment effects across diverse clinical problems. A 20% risk reduction may be highly important in one disease setting, but unimportant (or counteracted by drug toxicity) for a different disease or different outcome.

We estimated the Pearson correlation coefficients for the data presented in the Villar protocol as 0.76 (P < .001), which coincides with the estimates obtained in the Cappelleri protocol (r = 0.73, P < .001 for both the size and power approach). Thus, these protocols would suggest that large treatment benefits in large trials cluster with large treatment benefits found in respective meta-analyses, while small or no benefits in large trials are accompanied by a similar lack of efficacy in respective meta-analyses. In stark contrast (Figure 2), there was no correlation in the LeLorier protocol (r = 0.12, P = .5). The lack of correlation may be due to journal selection or to consideration of secondary end points. When only 1 primary end point was considered for each comparison, the estimated correlation in the LeLorier protocol became r = 0.50 (P = .07).

### Significance of the Difference in the Standardized Treatment Effects

The LeLorier and Cappelleri protocols used statistical tests to compare the standardized treatment effect (such as the logarithm of OR or risk ratio) in large tri-
Figure 2.—Plots of the odds ratio of large trials and the odds ratio of corresponding meta-analyses in the Villar protocol (excluding the large trial from each meta-analysis) and the LeLorier protocol. Each pair of large trial and corresponding meta-analysis is shown by an ellipse with dimensions proportional to the square root of the sample size of the large study and the meta-analysis. Primary end-point comparisons are shown by shaded ellipses and secondary end points by open ellipses. Note the strong correlation in the Villar protocol, the total lack of correlation in the LeLorier protocol when all comparisons are considered, but modest correlation when only primary end points are considered. The data have been reextracted from the cited studies. Three outlier comparisons in the Villar protocol are not shown. See JAMA for a comparable figure for the Cappelleri protocol.

Fixed and Random Effects and Variable Metrics

Another issue is how the pooling of trial results is performed in the meta-analysis and, when several large trials are available, among large trials. Fixed effects are more likely than random effects to reach formal statistical significance. Some meta-analyses reach formal statistical significance by fixed effects, but not by random effects; the converse is rare. Depending on whether the corresponding large trials have significant or nonsignificant results, there may be fewer or more discrepancies by fixed effects rather than by random effects when agreement between large trials and meta-analyses is judged on the level of statistical significance ($P < .05$ or $P > .05$). When agreement is based on whether the estimates are different beyond chance given their CIs, then disagreements tend to increase when fixed-effects calculations are used since they have more narrow CIs than random effects.

The importance of the method of assessing disagreements was illustrated in the Cappelleri protocol, where disagreements were double with fixed- vs random-effects calculations (the Villar and the LeLorier protocols presented fixed effects estimates). Finally, different treatment effect metrics were used in the different protocols (OR and risk ratio), but it is unlikely this would have affected the findings. The relative merits of the OR vs the risk ratio have been extensively debated, but for practical purposes, their use is equally appropriate in most situations.

Clinical Agreement and Disagreement

No protocol can offer a summary perspective of the clinical importance of all disagreements, since this is largely subjective, even if most important, to assess. In each different medical question where large trials and meta-analyses are compared to 5% expected by chance alone).
The strength of both clinical trials and of meta-analyses lies in their quantitative and systematic approach. Efforts to standardize the unbiased conduct and reporting of both randomized trials and meta-analyses should be enhanced.4,5,46 We have come a long way in generating fairly reliable evidence through controlled clinical research, but we are still exposed to the possibility of unexpected discrepancies and surprises. Both meta-analyses and randomized controlled trials should be scrutinized for biases and diversity.6 A major role for meta-analysis in the future would be to understand and predict discrepancies and plan clinical trials with meta-analyses in mind.1 Prospective advance planning of clinical trials with the explicit anticipation that they will be incorporated in a meta-analysis also needs to be encouraged,10,46 and we should view evidence from small and large trials and meta-analyses as offering a complementary evolving continuum.

References

©1998 American Medical Association. All rights reserved.