

# Measuring the Quality of Trials

## The Quality of Quality Scales

Jesse A. Berlin, ScD

Drummond Rennie, MD

PHYSICIANS SEEKING THE BEST INFORMATION ABOUT PARTICULAR interventions often turn to the results of meta-analyses. Meta-analyses, if done correctly according to explicit rules, will include all relevant studies that meet specified criteria, even those unpublished, to produce an unbiased estimate of the intervention's worth. If the quality of the component studies of a meta-analysis is poor, then a precise summary of those poor studies is unjustified. Since poor-quality studies sometimes produce systematically different results, for example, larger treatment effects,<sup>1,2</sup> a meta-analysis may not only be deceptively precise, but may yield misleading results. In an attempt to deal directly with issues of study quality, many meta-analyses of therapeutic issues restrict consideration to randomized controlled trials.

Experience has shown, however, that even randomized controlled trials sometimes show bias.<sup>1-3</sup> What has become something of an issue among meta-analysts, is how best to measure study quality and to account for it in performing meta-analyses. Even within randomized controlled trials, quality is an elusive metric. Jadad and colleagues<sup>4</sup> define quality as "the likelihood of the trial design to generate unbiased results" but as Verhagen et al<sup>5</sup> point out in an article that displays the difficulty in getting agreement on what constitutes quality, this only covers internal validity.<sup>5</sup> A complete definition of quality also should take into account the trial's external validity and its statistical analysis, as well as, perhaps, its ethical aspects.

Why would physicians be concerned about evaluating trial quality as part of a meta-analysis? Physicians would be most interested in knowing whether high-quality trials give different summary estimates of treatment effect than low-quality trials. If high-quality trials do give systematically different estimates of treatment effect, then physicians would want to draw clinical conclusions on the basis of high-quality not low-quality trials. This could be done using stratification on quality.

Systems for assessing quality are of 2 general kinds. There are those that simply assess the presence or absence in the report of a trial of a few key components, such as adequate

concealment of allocation.<sup>1,2</sup> Others use scales in which various items (for example, descriptions of randomization or of dropouts) are given numerical scores that are then totaled.

In this issue of THE JOURNAL, 2 complementary articles address these clinical concerns.<sup>6,7</sup> Jüni and colleagues<sup>6</sup> examine trials comparing low-molecular weight heparin with standard heparin for thromboprophylaxis in general surgery. The authors show that whether a trial is called high or low quality depends on which of the 25 scoring scales is used. Consequently, in a meta-analysis, summarizing the high-quality trials separately from the low-quality trials leads to drastically different clinical conclusions, depending again on which scale is used. The figure that Jüni and colleagues present is truly astounding. It shows that the difference in treatment effects between purportedly high-quality and low-quality trials changes drastically in magnitude and even may change in direction, depending on the choice of quality scale.

Despite seemingly dramatic differences between high-quality and low-quality scores, there were no statistically significant associations between individual quality scores and treatment effect. At worst, it seems that scales (with more statistical power) are useless. At best, these scales would show no significant relation to effect size. Particular features of studies, however, do relate to effect size.

If such evaluations, applied to other meta-analyses, are found to depend as heavily on the choice of quality scales as Jüni et al found in this meta-analysis, then readers cannot rely on quality scales to reach clinical conclusions. However, as Jüni et al point out, it is not surprising that the scales give such different results, as they focus on different aspects of trials. Some focus more on aspects of reporting of trials than on the design of the trials. Good quality reporting helps the reader evaluate the quality of the design. Even when certain aspects of trial design may be deficient methodologically, thorough reporting allows the reader to put the trial results in perspective.

In attempting to draw clinical conclusions, though, it is necessary to have trials in which the potential for bias has been reduced. If the results of Jüni et al are confirmed, it follows that scales should be abandoned, and that readers should concentrate on key components of design, rather than reporting.

**Author Affiliations:** School of Medicine, University of Pennsylvania, Philadelphia (Dr Berlin), and JAMA, Chicago, Ill (Dr Rennie).

**Corresponding Author and Reprints:** Drummond Rennie, MD, 515 N State St, Chicago, IL 60610.

See also pp 1054 and 1061.

One might argue that the underlying issue relevant to the clinical interpretation of meta-analyses is heterogeneity of trial results. When trials do not give a uniform estimate of treatment effectiveness, readers want to understand the sources of variability in the measured treatment effect. Particular aspects of trial design, such as blinding of the assessment of outcome, offer potential explanations for variability of treatment effects across trials. Heterogeneity of trial results may be due to a great many reasons. Such features as timing and dosage of medications and clinical characteristics of the populations being studied can also influence the true effectiveness of treatments. These features may vary among trials and should not be ignored when considering the variability of effects. They may not detract from the quality of individual trials but might affect the validity of pooling the results of different trials and hence the quality of meta-analyses.<sup>8,9</sup> In a given situation, Jüni et al point out, one might expect different specific aspects of design to matter with respect to the measured treatment effect. Where there is an outcome that requires subjective judgment to measure, such as assessing deep vein thrombosis, blinding of the outcome assessment may matter, whereas the reader would not expect blinding to matter when, for example, mortality is the outcome.

Jüni and colleagues are not the first to note that quality scales may not be the only useful approach to helping with the clinical interpretation of meta-analyses. Greenland<sup>10</sup> also argues for using specific aspects of study design. If one particular aspect of design is related to outcome, then using quality scores (which are at best arbitrary in their assignment of weights to responses) risks loss of information.

Beyond helping to interpret the clinical message in the meta-analysis, evaluation of the specific design components also seems more helpful than quality scores when designing subsequent trials. It seems less helpful to know that there is a need to design better quality trials, which we might take as a given, than to know that future trials should include specific features, such as blinded assessments of outcome.

Relationships between specific aspects of study design and study outcomes are also explored in this issue of THE JOURNAL by Lijmer and colleagues.<sup>7</sup> It is one thing to hypothesize sources of bias in trials and quite another to provide empirical evidence that, in practice, such bias does exist and can distort clinical understanding of diagnostic test properties. The article by Lijmer et al is an excellent example of using the theory from the literature on bias in studies of diagnostic testing to guide an analysis that aims to quantify bias.

Using an extension of methods for fitting summary receiver operating characteristic curves across studies<sup>11,12</sup> these authors found 2 particular aspects of study design that were strongly associated with accuracy of diagnostic tests. One of these was the use of the case-control design, a design in which test performance among patients known to

have disease because of clear clinical or other evidence, is compared with performance among subjects known to be free of disease. In other words, the patients included in the study were at the extremes of the spectrum of disease (either healthy or sick) in contrast with a more typical clinical situation, in which the physician would be more likely to encounter patients with a continuum of disease severity. This type of design led to clear overstatements of accuracy in the examples studied by Lijmer et al.<sup>7</sup> Although this conclusion was based on a small number of studies, it seems fairly consistent with the expectation based on prior literature.<sup>13-17</sup> Similarly, use of different reference tests for those with positive and negative index tests also led to overstatement of accuracy.

Both of the findings about study design should suggest that future studies should avoid such designs. Unfortunately, when the definitive diagnostic test is invasive or otherwise dangerous, its use in conducting research raises ethical problems. The solution to this dilemma is not at all clear.

Two other findings by Lijmer and colleagues raise different sorts of issues. The authors found that studies that provide inadequate descriptions of either the test performed, or the population studied, also tend on average to report higher accuracy than better-described studies. This finding reinforces our earlier concern about distinguishing between poor quality of reporting and poor quality of design. It seems unlikely that mere poor reporting is an adequate explanation of the apparent bias in test accuracy. More plausibly, either clearer reporting would reveal deficiencies in study design that might underlie the bias, or poorly reported studies have other unmeasured design flaws more often than clearly reported studies.

Most of those who have tried to assess the quality of individual trials as revealed by their published reports have concluded that quality is often low.<sup>2,18</sup> This is disturbing and one reason for the CONSORT initiative.<sup>19,20</sup> While it is clear that some of the CONSORT reporting requirements will need to be revised,<sup>21,22</sup> there seems little doubt that in journals that have adopted it, CONSORT has increased the completeness of reporting (David Moher, oral communication, May 1999). Now Lijmer et al are assisting those who wish to analyze diagnostic tests to increase the trustworthiness of their studies. The 2 studies in this issue of THE JOURNAL show that, given the clinical consequences, everyone has a stake in the quality of assessing quality.

#### REFERENCES

- Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. 1995;273:408-412.
- Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet*. 1998;352:609-613.
- Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignments in controlled clinical trials. *N Engl J Med*. 1983;309:1358-1361.
- Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials*. 1995;17:1-12.
- Verhagen AP, de Vet HCV, de Bie RA, et al. The Delphi list: a criteria list for

quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epidemiol.* 1998;12:1235-1241.

6. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA.* 1999;282:1054-1060.

7. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 1999;282:1061-1066.

8. Greenland S. A critical look at some popular meta-analytic methods. *Am J Epidemiol.* 1994;140:290-296.

9. Berlin JA, Antman EM. Advantages and limitations of meta-analytic regressions of clinical trials data. *Online J Curr Clin Trials* [serial online]. June 4, 1994; doc 134.

10. Greenland S. Quality scores are useless and potentially misleading. *Am J Epidemiol.* 1994;140:300-301.

11. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med.* 1993;12:1293-1316.

12. Irwig LI, Tosteson ANA, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med.* 1994;120:667-676.

13. Reid MC, Lachs MS, Feinstein AR. Use of methodologic standards in diagnostic test research: getting better but still not good. *JAMA.* 1995;274:645-651.

14. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature, III: how to use an article about a diagnostic test, A: are the results of the study valid? *JAMA.* 1994;271:389-391.

15. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature, III: how to use an article about a diagnostic test, B: what are the results and will they help me in caring for my patients? *JAMA.* 1994;271:703-707.

16. Greenhalgh T. How to read a paper: papers that report diagnostic or screening tests. *BMJ.* 1997;315:540-543.

17. Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. *J Gen Intern Med.* 1989;4:288-295.

18. Ioannides JPA, Lau J. Can quality of clinical trials and meta-analyses be quantified? *Lancet.* 1998;352:590.

19. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT Statement. *JAMA.* 1996;276:637-639.

20. Rennie D. How to report randomized controlled trials. *JAMA.* 1996;276:649.

21. Meinert CLM. Beyond CONSORT: need for improved reporting standards for clinical trials. *JAMA.* 1998;279:1487-1489.

22. Moher D. CONSORT: an evolving tool to help improve the quality of reports of randomized controlled trials. *JAMA.* 1998;279:1489-1491.

# Fourth International Congress on Peer Review in Biomedical Publication: Call for Research

Drummond Rennie, MD

Annette Flanagin, RN, MA

Richard Smith, MD

Jane Smith, MSc

**T**HE INCREASING SUCCESS OF THE FIRST 3 PEER REVIEW CONGRESSES,<sup>1-3</sup> held at 4-year intervals, encourages us to hold a fourth congress in 2001. The Fourth International Congress on Peer Review in Biomedical Publication will be held September 13-16, 2001, in Barcelona, Spain.

This congress, organized by *JAMA* and the *BMJ* Publishing Group, will feature 3 days of presentations of original research. As before, we urge scientists, editors, and readers who are interested in the processes by which science is published to get going on their research. Topics of interest include the mechanisms of peer review and editorial decision making and evaluations of their validity and practicality; online and Web-based peer review, publication, and pre-publication posting and release of information; quality assurance for reviewers and editors; authorship, contributorship, and responsibility for published material; breakdowns, weaknesses, and biases; peer review of grant proposals; conflicts of interest; scientific misconduct; the

**Author Affiliations:** Dr Rennie is Deputy Editor (West) and Ms Flanagin is Managing Senior Editor, *JAMA*; Dr Smith is Editor and Ms Smith is Deputy Editor, *BMJ*. **Corresponding Author and Reprints:** Annette Flanagin, RN, MA, *JAMA*, 515 N State St, Chicago, IL 60610 (email:annette\_flanagin@ama-assn.org).

economics of peer review and scientific publication; methods for improving the quality, efficiency, and equitable distribution of biomedical information; methods for measuring the quality of print and online information; interactive digital systems and other new technologies that affect the dissemination of biomedical information; and the future of scientific publication.

Abstracts on any aspect of editorial peer review and its role in scientific publication and information exchange will be considered. Abstracts that summarize new research and findings will be given priority. Deadline for submission of abstracts is January 15, 2001. Instructions for preparing and submitting abstracts are available on the Peer Review Congress Web site at <http://www.ama-assn.org/peer> or by contacting Jennifer Reiling, *JAMA*, 515 N State St, Chicago, IL 60610 USA; telephone: 312-464-5108; fax: 312-464-5824; e-mail: jennifer\_reiling@ama-assn.org.

As with the previous congresses, our aim is to improve the quality and credibility of biomedical information and to help advance the efficiency, effectiveness, and equitability of information dissemination throughout the world.

## REFERENCES

1. Guarding the guardians: research on editorial peer review: selected proceedings from the First International Congress on Peer Review in Biomedical Publication. *JAMA.* 1990;263(theme issue):1317-1441.
2. The Second International Congress on Peer Review in Biomedical Publication. *JAMA.* 1994;272(theme issue):91-173.
3. The Third International Congress on Peer Review in Biomedical Publication. *JAMA.* 1998;280(theme issue):203-306.