



JUDGING THE QUALITY OF MEDICAL LITERATURE*

By Teresa L. Rogstad, MPH

Medical Research Analyst, Hayes, Inc., Lansdale, PA

ABSTRACT

Medical writers are medical literature consumers. They need to be able to evaluate the quality of the articles they use as information sources or choose to cite in their own writing. Writers without training in research design may intuitively recognize well-done reviews and clinical studies. However, a more deliberate consideration of certain criteria will permit the most efficient use of both reviews and clinical studies. A well-constructed systematic review addresses a focused study question or questions, specifies the process used to identify relevant clinical studies, critically evaluates those studies, synthesizes the findings, and forms conclusions. Individual clinical studies may be judged according to their methodologic strength and application usefulness. Type of research design, study conduct, sample size, and the manner in which data are reported determine study strength. Generalizability, realistic selection of study participants, and patient-centered outcomes contribute to the usefulness of a study.

Medical writers are medical literature consumers. They need to be able to evaluate the quality of the articles they use as information sources or choose to cite in their own writing. Writers without training in research design may intuitively recognize good articles. However, a more deliberate consideration of certain criteria will permit the most efficient use of both reviews and clinical studies. The goal of this article

is to provide an overview of those criteria. The discussion will begin with a description of systematic reviews and their usefulness to medical writers. The majority of this article will review principles involved in the critical appraisal of clinical studies.

RECOGNIZING HIGH-QUALITY REVIEWS

Almost everyone who makes use of published medical literature has occasion to find a good review article. For purposes of obtaining background information or determining typical practice patterns, a well-written narrative review article may suffice. However, for a comprehensive overview of the clinical evidence pertaining to a particular issue, a systematic review is more likely to be useful. In a systematic review, there is a methodical search for and synthesis of the results of clinical research. Compared with narrative reviews, systematic reviews tend to be more comprehensive in terms of cited research findings, are less likely to be biased, and are more likely to be organized around explicit clinical questions. A narrative review primarily reflects the knowledge and opinions of an expert or group of experts, whereas a systematic review attempts to discover something new through a methodical analysis of published research evidence.

To be considered “systematic,” a review should include the following elements¹⁻³:

- **Focused study question(s):** A systematic review begins with a definition of what is important to know and then continues with a search for the answer. Examples of study questions include the follow-

ing: Is core decompression more effective than pain medication in delaying hip replacement? Is pancreas transplantation effective in preventing or reversing secondary complications of diabetes?

- **A specific search strategy:** Search refers to literature search, ie, a review of databases such as MEDLINE, EMBASE, or the Cochrane Library. The search should be systematic, and the strategy should be defined in the review article. Elements of such a strategy include at a minimum the particular databases searched, publication dates included, and search terms.
- **Specified criteria for article selection:** Further criteria should be applied to the selection of individual articles from the results of the literature search. Study design (eg, only randomized controlled trials), size of the study sample, and follow-up of a minimum duration are examples of selection criteria. These criteria should allow selection of articles most likely to provide the strongest and most applicable evidence, given the study question(s).
- **Critical appraisal of studies, including formal quality assessment:** This step of a systematic review applies quality criteria to the selected studies (see the following section). A good systematic review not only summarizes the reported findings of the selected studies but also provides comments on the strength of those studies, the strength of the relationship between those studies and the focused question(s) of the review, and any other qualifiers that might affect interpretation.

*This article is based on the content of the AMWA workshop (#99) of the same title.

Some authors of systematic reviews use formal checklists to assess the quality of individual studies.

- **Synthesis:** A good systematic review culminates in a synthesis of the evidence provided by the individual clinical studies, taking into account the critical appraisal of those studies. Synthesis may be purely qualitative, or it may involve statistical meta-analysis, in which data from multiple studies are pooled to derive more precise estimates of a particular treatment outcome or diagnostic accuracy. The synthesis leads to a stated conclusion.

A systematic review may stand alone. It may also be combined with other information and considerations to yield clinical policy (practice guidelines), reimbursement policy, or public health policy.^{2,4-6} For a general medical writer's purposes, a systematic review can be useful for either providing a well-reasoned answer to a particular issue or simply identifying the best clinical studies on a topic.

RECOGNIZING HIGH-QUALITY CLINICAL STUDIES

Various medical writing tasks require the interpretation of individual clinical studies. Some degree of critical appraisal of study quality is more likely to lead to selection of the best studies and to a more accurate representation of their findings. The quality of evidence provided by a study is derived from its strength as well as its usefulness. The following discussion does not provide a complete guide to critical appraisal of clinical research but illustrates key principles. Additional information and quality checklists are available online and in the literature.^{1, 3, 4, 7-12}

Study Strength

The terms *strength of evidence* or *study strength* refer primarily to methodologic strength. It refers to measures taken by investigators to enhance the

internal validity of the study, which is another way of saying that bias is minimized. The strength of a study is usually assessed in terms of research design, study conduct, sample size, and reporting and analysis. Research design typically determines the *level of evidence*, which might then be upgraded or downgraded based on the other aspects of the study or study article.

Research Design

A hierarchy of study design guides the first step in assessing the strength of an individual study. Epidemiology textbooks are good resources for detailed discussions of specific study designs. Evidence-grading schemes typically use some variation of the following simple categories, at least for studies of treatment interventions.^{1,4,13-17} These categories are listed in order from strongest to weakest:

- Randomized controlled/comparative trials
- Nonrandomized controlled/comparative studies
- Uncontrolled/noncomparative studies
- Expert opinion, case reports

Some systems include meta-analyses (of randomized controlled trials) in the same category as randomized controlled trials. Other systems do not include meta-analyses at all because they do not provide primary evidence. In a randomized controlled trial, patients are assigned either to the active treatment group or to a control group that does not receive this treatment. A computerized algorithm makes a random assignment each time a patient is enrolled. The result is that the 2 groups (treatment and control) are as similar as possible. They are unlikely to differ in factors that might bias results by differentially affecting treatment response in the 2 groups. Such factors are called confounders. An example of a confounder would be age in a trial in which the patients undergoing usual conser-

vative treatment (the control group) were on average younger than patients undergoing a new surgical procedure (the interventional group). A lower incidence of future cardiovascular events in the control group could be partially attributable to the age difference. With successful randomization, any observed difference in outcome should be due solely to the difference in interventions and not because of pre-existing patient differences. A randomized controlled trial compares the active intervention of interest either with a placebo (no treatment) or with standard treatment. Head-to-head or comparator trials, in which 2 nonstandard alternatives are compared, may also be randomized. Randomized trials are by their very nature prospective in design, with patients enrolled according to a specified study protocol and data collected to answer study questions. A prospective study of any design is one in which outcomes are not yet known.

Unlike randomized trials, nonrandomized controlled/comparative studies are subject to selection bias because factors that affect intervention assignment may also be related to outcome. For example, more severely ill patients might be more likely to volunteer for an experimental treatment because they have more to gain, whereas less severely ill patients might prefer to remain with standard treatment. Nonrandomized studies may be prospective or retrospective. Retrospective studies analyze data pertaining to patients treated in the past; the outcome of treatment is already known. Such studies are subject to information bias because of the limited availability, accuracy, and completeness of patient charts, or by previous collection of data without research purposes in mind. Another variation is to prospectively assess a group of patients and then make comparisons with a historical control group, that is, a group treated earlier in time. This design carries the limitations of retrospective data analysis

and may preclude the observance of identical selection criteria and treatment protocols for the 2 groups (see the Study Conduct section).

Uncontrolled, noncomparative studies include longitudinal studies, case series, and database analyses that do not compare 2 groups of patients. Longitudinal studies are prospective by definition. Case series and database reviews are generally considered retrospective studies although authors sometimes state that data collection was prospectively defined.

Another way of categorizing research designs is to group them as experimental/interventional or observational. Experimental studies represent deliberate intervention, or treatment assignment, on the part of the investigator. The term usually brings to mind randomized trials, but nonrandomized methods of treatment assignment characterize some experimental studies. In observational studies, the investigators do not make treatment assignments. Observational studies may involve natural control or comparison groups (as in cohort studies), comparisons of current patients with historical controls, case-control studies, cross-sectional studies (patient-level data), and ecologic/correlational studies (group-level data). Other observational study types are database analyses, which may or may not involve comparisons, and case series. Experimental studies are considered to be methodologically stronger than observational studies but may not be as useful (see the Study Usefulness section).

The lowest level of clinical evidence comprises certain sources of information rather than actual study designs. This level includes expert opinion and case reports. The opinions of experts, even of experts with vast experience, are subject to bias due to knowledge gaps, nonrepresentative patient populations, and the limitations of human perception and memory. Furthermore, experts in the field may tend to be advocates

of new technology because they work at institutions most likely to be early adopters. Nevertheless, in the absence of clinical trial data, expert opinion can be very useful. Case reports, published descriptions of a single case or a small number of cases, may suggest avenues of research but do not represent systematically derived evidence. Practice guidelines often must rely on expert opinion to address some issues, but health technology assessments and systematic reviews generally exclude expert opinion and case reports. (Practice guidelines recommend approaches to multiple aspects of a particular disease or clinical problem; health technology assessments evaluate the safety and effectiveness of devices, drugs, procedures, or tests. A practice guideline may make use of one or more previously written health technology assessments.)

Study Conduct

Apart from design category, many choices made by investigators in the conduct of a trial can affect study quality. In evidence-grading schemes, factors related to study conduct might positively or negatively modify a ranking made solely because of study design. For randomized trials, blinding is an important quality differentiator. Blinding means that the persons involved in a trial do not know the identity of the interventions being delivered to specific patients until after the completion of data collection. If patients, clinicians administering the interventions, or other personnel involved in data collection and analysis are aware of treatment assignment and if there is any subjective element to the reporting of symptoms or evaluation of outcomes, then the results might partially reflect expectations associated with the newer treatment. Thus, bias would be introduced and some of the benefits of randomized treatment assignment would be lost. Additionally, if patients knew that they were not in the experi-

mental/treatment group, they might be more likely to drop out of the trial, resulting in large losses to follow-up. In single-blind trials, the study participants are unaware of which intervention they are receiving. In double-blind trials, both the patient and the evaluators are unaware of the intervention received.

In nonrandomized controlled/comparison studies, it is important that patient groups be made as similar as possible. Researchers can specify the same inclusion/exclusion criteria for patient enrollment in all groups. Such criteria are usually related to factors such as medical history, comorbidities, age, sex, and disease severity. Investigators may go a step further and select a control group or individual control patients on the basis of characteristics matched to the specific characteristics of patients already selected for the treatment group. Bias may also be introduced by differences in treatment settings or the timing and manner of outcome assessment. In any controlled or comparison study, regardless of randomization, the study protocol should be well defined so that there are minimal treatment differences between groups except for those related to the intervention of interest. For example, in a study comparing extracorporeal shock wave treatment for tennis elbow with a sham treatment (a form of placebo), supplemental corticosteroids should be either prohibited or allowed in both groups.

Lastly, a less-than-adequate follow-up interval, losses to follow-up, or both may invalidate the results of an otherwise strong study. Follow-up must be long enough to allow the outcomes being reported to manifest themselves. A study focusing on short-term adverse reactions may need only a brief follow-up interval. A study of techniques used in fracture repair requires perhaps only a few months of follow-up. However, a study reporting the impact of a cholesterol-lowering drug on cardiovascular events

requires long-term follow-up of several years. As the follow-up interval lengthens, the possibility of loss of patients to follow-up increases. Over time, some patients may discontinue their assigned treatment or not return for follow-up visits. A difference in follow-up rates between comparison groups suggests biased results. Even in an uncontrolled case series, if the reasons for not returning are related to how well or how poorly patients fared following treatment, results can be misleading. The careful reader will look to see whether measures have been taken to minimize loss to follow-up or to compensate for losses by such tactics as telephone interviews.

Sample Size

There is no magic number when it comes to adequate sample size. In larger samples, the results of the study are more likely to be consistent with results that would be observed if the whole population of interest were studied. In other words, large samples reduce the likelihood of sampling error. In situations in which the magnitude of expected improvement or the magnitude of the expected difference between 2 groups is small, a larger sample size is necessary for observed changes or differences to be statistically significant. Thus, one mark of a carefully planned study is the report of *power calculations*, in which minimum sample size is determined ahead of time and is based on expectations of differences of a specified magnitude. These expectations may be based on what the authors consider to be a clinically important effect.

Reporting and Analysis of Results

Investigators have not added to the overall body of clinical evidence if they have simply collected data. Quality in the dimension of reporting and analysis means that data were analyzed in a manner appropriate to the clinical issues and to the nature of the data. It also means

that enough data are reported for the reader to judge the authors' conclusions. The statistical significance and the variability of key results should be reported. Statistical significance indicates the probability (*P* value) that change or the difference between groups occurred by chance alone. Variability refers to the level of precision in observed results as expressed by a confidence interval, standard deviation, standard error of the mean, or simply by a range of observed values. Even readers who are untrained in statistics can look to see that these issues were considered. (For more

information on evaluating statistics, see the series of articles by Tom Lang published in the *AMWA Journal*.) Testing for statistical significance should not preclude reporting clinical significance, also referred to as clinical importance. For example, a pain treatment may result in a statistically significant 1-point decrease on a 10-point pain scale. Is that enough improvement to make a difference in patient well being or functional abilities? It may not be possible for a medical writer unfamiliar with the field to make these judgments, but data showing the magnitude of effect

Reporting Standards

For medical writers who work with authors to prepare articles for publication, reporting standards are very important tools. Reporting standards were not developed to guide critical appraisal of already published literature, but to ensure that biomedical articles being written for publication include the information necessary to permit critical appraisal. These standards were derived from principles of good research and analysis.^{1,2} Widely recognized reporting guidelines include the following.

- PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)
- CONSORT Statement (Consolidated Standards of Reporting Trials [for randomized controlled trials])
- STARD (Standards for Reporting of Diagnostic Accuracy)
- STROBE (Strengthening the Reporting of Observational studies in Epidemiology)
- MOOSE (Meta-Analysis Of Observational Studies in Epidemiology)
- TREND (Transparent Reporting of Evaluations with Nonrandomized Designs [focuses on behavioral and public health interventions])

These checklists and other similar resources are available through an online clearinghouse managed by the EQUATOR Network.³

1. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Fam Pract.* 2004;21:4-10.
2. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med.* 2001;134:663-694.
3. EQUATOR Network. Resource Centre: Library of health research reporting. Available at www.equator-network.org/index.aspx?o=1032#what. Accessed October 12, 2009.

and some comment on the part of the authors about clinical importance are the marks of good analysis.

Analytic techniques are important ways to compensate for deficiencies of study design or unavoidable problems. In nonrandomized controlled/comparative studies, various statistical techniques can be used to control for known confounders so that results are adjusted for baseline differences between groups and bias is minimized. There are also techniques for calculating results in multiple ways to explore the possible effects of high losses to follow-up or the effect of patients' unplanned crossover from one treatment to another.

Study Usefulness

A study may or may not be applicable to all clinical and policy situations, even if well designed and conducted in an appropriate manner. Generalizability, also referred to as applicability or external validity, is a key concept of usefulness. Can the results obtained in the particular setting and for the particular patients represented by the study be generalized to other settings and patients? This area is often deficient in randomized trials. The very things that help minimize bias and aid clear interpretation—tightly controlled treatment and strict monitoring protocols, highly trained clinical staff, carefully selected patients—give rise to the question of whether the same results would be obtained in routine practice settings and among the patient populations typically seen in those settings. Furthermore, because of their expense, randomized trials usually do not have long follow-up periods and thus may not shed light on outcomes such as survival or long-term safety. Randomized trials are generally designed to evaluate efficacy—how well the treatment or diagnostic intervention works in a controlled setting. In the early stages of development for a new intervention, efficacy and safety are the most important issues.

Eventually, studies that address effectiveness—how well the intervention works in typical practice settings—are the more useful studies. There are randomized effectiveness trials, also called pragmatic trials; however, the primary source of effectiveness data is observational studies. Researchers are developing new observational study designs and statistical tools to serve the growing demand for real-world effectiveness data.^{18, 19}

The manner in which a study sample is selected also affects generalizability. Although randomization minimizes bias by making treatment groups or treatment and control groups similar, it is possible for the overall study group that undergoes randomization to have been selected in a manner that is not systematic and is thus not representative. Enrolling every consecutive and eligible patient or selecting a random sample from the eligible population would be good ways to achieve sufficient representation in controlled/comparative trials and uncontrolled studies.

Another drawback to a study's usefulness is the type of outcome measured. Intermediate outcomes are less useful ultimately than health-related or patient-centered outcomes. For example, the effect of a surgical technique on range of motion is less meaningful than whether it helped the patient return to playing tennis. The effect of antihypertensive medication on blood pressure is crucial, but whether use of the medications prevents heart attacks and strokes is even more important. Results expressed in terms of quality-adjusted life-years or healthy life-years are especially meaningful from the patient's viewpoint. Lastly, cost-effectiveness may be a useful outcome measure from a payer or policymaker perspective.

Other issues affect usefulness. Depending on the developmental stage of the intervention, trials comparing the intervention with another relatively new alternative may be more useful than those comparing

the intervention with a placebo or with standard treatment. If effectiveness has been established, then studies analyzing long-term safety issues may be needed. Studies that attempt to answer remaining questions about the use of the intervention in certain high-risk groups such as the elderly may be the most useful for interventions that have already been well studied in general populations.

STUDIES OF NONTHERAPEUTIC INTERVENTIONS

Quality criteria for studies of diagnostic, prognostic, and screening methods are not as well developed or as easy to comprehend as those for therapeutic studies. The principles discussed so far are most easily applied to therapeutic studies, ie, studies that evaluate treatments. These principles can also be applied to studies evaluating diagnostic/prognostic tests when the studies are designed to measure an outcome, eg, a change in treatment plan or improvement in survival.²⁰ Such assessments of clinical impact imply the comparison of patients who are treated according to results of the new or unproven test with patients who are treated according to standard criteria. However, most nontherapeutic studies stop short of evaluating clinical outcomes. At best, they calculate sensitivity and specificity by comparing test results with results of a so-called gold standard (reference standard) or with surgical/pathologic confirmation. A typical disease mix in the tested population, a reasonable source of reference (normal) values, and blinded evaluation of test results improve the validity of nontherapeutic studies, whether they are assessing outcomes or accuracy. Examples of even less informative studies are those that simply explore statistical associations between laboratory test results and known disease or describe subjective evaluation of image quality for an imaging technique.

CONCLUSION

Many of the principles that define the quality of medical literature are within the realm of common sense and will not be surprising to medical writers, even if some of the terms are new. Moreover, effective use of the clinical literature does not necessarily require the detailed and technical critique that this review might imply. It is my hope, however, that the reader can now more quickly discriminate between the best and the not-so-good.

Acknowledgments

I thank my colleagues at Hayes, Inc.—Anita Rihal, MPH, and Sharon Selman, PhD—as well as Katherine L. Kraines (freelance writer/editor) for their feedback during preparation of this article.

Author disclosure: The author notes the following commercial relationship: full-time employment at Hayes, Inc., a producer of health technology assessments.

References

1. EBM Tools. Critical Appraisal and Evidence Grading. *Centre for Evidence-Based Medicine*. Available at www.cebm.net/?o=1023. Accessed January 22, 2009.
2. IOM Reports. Knowing What Works in Health Care: A Roadmap for the Nation. Systematic Review: The Central Link Between Evidence and Clinical Decision Making. March 21, 2008. Available at www.iom.edu/CMS/3809/34261/50718/52680.aspx. Accessed January 22, 2009.
3. West S, King V, Carey T, et al. *Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47*. Rockville, MD: Agency for Healthcare Research and Quality; 2002. AHRQ publication 02-E016. 4. American Academy of Neurology. Practice Guidelines. Clinical Practice Guideline Process Manual, 2004 Edition. Available at www.aan.com/globals/axon/assets/3749.pdf. Accessed January 22, 2009.
4. Practice Guidelines. Clinical Practice Guideline Process Manual. 2004 Edition. American Academy of Neurology (AAN) Web site. Available at www.aan.com/globals/axon/assets/3749.pdf. Accessed January 22, 2009.
5. Agency for Healthcare Research and Quality. Clinical Information. Technology Assessments. Available at www.ahrq.gov/clinic/techix.htm. Accessed January 22, 2009.
6. U.S. Preventive Services Task Force. About the Task Force. Methods and Processes. Update on Methods Estimating Certainty and Magnitude of Net Benefit. Available at www.ahrq.gov/clinic/uspstmeth.htm. Accessed January 22, 2009.
7. Training resources. Cochrane Handbook for Systematic Reviews of Interventions. Available at www.cochrane-handbook.org/. Accessed January 22, 2009.
8. Coleman BD, Khan KM, Maffulli N, Cook JL, Wark JD. Studies of surgical outcome after patellar tendinopathy: clinical significance of methodological deficiencies and guidelines for future studies. Victorian Institute of Sport Tendon Study Group. *Scand J Med Sci Sports*. 2000;10:2-11.
9. Deville WL, Buntinx F, Bouter LM, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol*. 2002;2:9.
10. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials*. 1996;17:1-12.
11. Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.0.2 [updated September 2009]. The Cochrane Collaboration, 2008. Available at www.cochrane-handbook.org. Accessed January 22, 2009.
12. Cochrane Diagnostic Test Accuracy Working Group. Handbook for DTA Reviews. Available at <http://srdta.cochrane.org/en/authors.html>. Accessed January 22, 2009.
13. GRADE Working Group. Available at: <http://www.gradeworkinggroup.org/>. Accessed January 22, 2009.
14. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. *BMJ*. 2004;328:1490.
15. General Methodological Principles of Study Design. In: Appendix A, March 19, 2008 Decision Memo for Autologous Blood Derived Products for Chronic Non-Healing Wounds (CAG-00190R2). Available at <https://www.cms.hhs.gov/mcd/viewdecisionmemo.asp?id=208>. Accessed January 22, 2009.
16. Instructions to Authors. *The Journal of Bone and Joint Surgery (JBJS)—American Volume*. 2009. Available at www2.ejbs.org/misc/instrux.dtl#levels. Accessed January 22, 2009.
17. Quality and Science. Methodology Manual for ACC/AHA Guideline Writing Committees (Step Six: Assign Classification of Recommendations and Level of Evidence). 2006. Available at www.acc.org/qualityandscience/clinical/manual/pdfs/methodology.pdf. Accessed January 22, 2009.
18. Institute of Medicine. Public Meeting 4—Redesigning the clinical effectiveness research paradigm: innovation and practice-based approaches. Available at www.iom.edu/CMS/28312/RT-EBM/46337.aspx. Accessed January 22, 2009.
19. Garrison LP, Jr., Neumann PJ, Erickson P, Marshall D, Mullins CD. Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force report. *Value Health*. Sep-Oct 2007;10(5):326-335.
20. Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ*. 2008;336:1106-1110.

IN THE NEXT ISSUE...

- Continued coverage of the 2009 conference, with more session summaries
- Swanberg Address
- Details about the restructuring of AMWA's educational program

