Agency for Healthcare Research and Quality • www.ahrq.gov

Evidence Report/Technology Assessment

Number 47

Systems to Rate the Strength of Scientific Evidence

Summary

Introduction

Health care decisions are increasingly being made on research-based evidence rather than on expert opinion or clinical experience alone. Systematic reviews represent a rigorous method of compiling scientific evidence to answer questions regarding health care issues of treatment, diagnosis, or preventive services. Traditional opinion-based narrative reviews and systematic reviews differ in several ways. Systematic reviews (and evidence-based technology assessments) attempt to minimize bias by the comprehensiveness and reproducibility of the search for and selection of articles for review. They also typically assess the methodologic quality of the included studies—i.e., how well the study was designed, conducted, and analyzed—and evaluate the overall strength of that body of evidence. Thus, systematic reviews and technology assessments increasingly form the basis for making individual and policy-level health care decisions.

Throughout the 1990s and into the 21st century, the Agency for Healthcare Research and Quality (AHRQ) has been the foremost federal agency providing research support and policy guidance in health services research. In this role, it gives particular emphasis to quality of care, clinical practice guidelines, and evidence-based practice—for instance through its Evidence-based Practice Center (EPC) program. Through this program and a group of 12 EPCs in North America, AHRQ seeks to advance the field's understanding of how best to ensure that reviews of the clinical or related literature are scientifically and clinically robust.

The Healthcare Research and Quality Act of 1999, Part B, Title IX, Section 911(a) mandates that AHRQ, in collaboration with experts from the public and private sectors, identify methods or systems to assess health care research results, particularly "methods or systems to rate the strength of the scientific evidence underlying health care practice, recommendations in the research literature, and technology assessments." AHRQ also is directed to make such methods or systems widely available.

AHRQ commissioned the Research Triangle Institute–University of North Carolina EPC to undertake a study to produce the required report, drawing on earlier work from the RTI–UNC EPC in this area.[1] The study also advances AHRQ's mission to support research that will improve the outcomes and quality of health care through research and dissemination of research results to all interested parties in the public and private sectors both in the United States and elsewhere.

The overarching goals of this project were to describe systems to rate the strength of scientific evidence, including evaluating the quality of individual articles that make up a body of evidence on a specific scientific question in health care, and to provide some guidance as to "best practices" in this field today. Critical to this discussion is the definition of quality. "Methodologic quality" has been defined as "the extent to which all aspects of a study's design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error." (Ref. 1, p. 472) For purposes of this study, the authors hold quality to be the extent to which a study's design, conduct,

and analysis have minimized selection, measurement, and confounding biases, with their assessment of study quality systems reflecting this definition.

The authors do acknowledge that quality varies depending on the instrument used for its measurement. In a study using 25 different scales to assess the quality of 17 trials comparing low molecular weight heparin with standard heparin to prevent post-operative thrombosis, Juni and colleagues reported that studies considered to be of high quality using one scale were deemed low quality on another scale.[2] Consequently, when using study quality as an inclusion criterion for meta-analyses, summary relative risks for thrombosis depended on which scale was used to assess quality. The end result is that variable quality in efficacy or effectiveness studies may lead to conflicting results that affect analyst's or decisionmakers' confidence about findings from systematic reviews or technology.

The remainder of this summary briefly describes the methods used to accomplish these goals and provides the results of the authors' analysis of relevant systems and instruments identified through literature searches and other sources. They present a selected set of systems that they believe are ones that clinicians, policymakers, and researchers can use with reasonable confidence for these purposes, giving particular attention to systematic reviews, randomized controlled trials (RCTs), observational studies, and studies of diagnostic tests. Finally, they discuss the limitations of this work and of evaluating the strength of the practice evidence for systematic reviews and technology assessments and offer suggestions for future research. The authors do not examine issues related to clinical practice guideline development or assigning grades or ratings to formal guideline recommendations.

Methods

To identify published research related to rating the quality of studies and the overall strength of evidence, the authors conducted two extensive literature searches and sought further information from existing bibliographies, members of a technical expert panel, and other sources. They then developed and completed descriptive tables—hereafter "grids"—that enabled them to compare and characterize existing systems. These grids focus on important domains and elements that the authors concluded any acceptable instrument for these purposes ought to cover. These elements reflect steps in research design, conduct, or analysis that have been shown through empirical work to protect against bias or other problems in such investigations or that are long-accepted practices in epidemiology and related research fields. They assessed systems against domains and assigned scores of fully met (Yes), partially met (Partial), or not met (No).

Then, drawing on the results of their analysis, the authors identified existing quality rating scales or checklists that in their view can be used in the production of systematic evidence reviews and technology assessments and laid out the reasons for highlighting these specific instruments. An earlier version of the entire report was subjected to extensive external peer review by experts in the field and AHRQ staff, and the authors revised that draft as part of the steps to produce this report.

Results

Data Collection

The authors reviewed the titles and abstracts for a total of 1,602 publications for this project. From this set, they retained 109 sources that dealt with systems (i.e., scales, checklists, or other types of instruments or guidance documents) pertinent to rating the quality of individual systematic reviews, RCTs, observational studies, or investigations of diagnostic tests, or with systems for grading the strength of bodies of evidence. In addition, they reviewed 12 reports from various AHRQ-supported EPCs. In all, the authors considered 121 systems as the basis for this report.

Specifically, they assessed 20 systems relating to systematic reviews, 49 systems for RCTs, 19 for observational studies, and 18 for diagnostic test studies. For final evaluative purposes, the authors focused on scales and checklists. In addition, they reviewed 40 systems that addressed grading the strength of a body of evidence (34 systems identified from their searches and prior research and 6 from various EPCs). The systems reviewed totals more than 121 because several were reviewed for more than one grid.

Systems for Rating the Quality of Individual Articles

Important Evaluation Domains and Elements

For evaluating systems related to rating the quality of individual articles, the authors defined important domains and elements for four types of studies. Boxes A and B list the domains and elements used in this work, highlighting (in *italics*) those domains they regarded as critical for a scale or checklist to cover before they could identify a given system as likely to be acceptable for use today.

Systematic Reviews

Of the 20 systems concerned with systematic reviews or meta-analyses, the authors categorized one as a scale[3] and 10 as checklists.[4-14] The remainder are considered guidance documents.[15-23]

To arrive at a set of high-performing scales or checklists pertaining to systematic reviews, the authors took account of seven key domains (see Box A): study question, search strategy, inclusion and exclusion criteria, data abstraction, study quality and validity, data synthesis and analysis, and funding or sponsorship. One checklist fully addressed all seven domains.[7] A second checklist also addressed all seven domains but merited only a "Partial" score for study question and study quality.[8] Two additional checklists [6,12] and the one scale[23] addressed six of the seven domains.

Box A. Important Domains and Elements for Systems to Rate Quality of Individual Articles

Systematic Reviews

- *Study question*
- *Search strategy*
- *Inclusion and exclusion criteria*
- Interventions
- Outcomes
- *Data extraction*
- Study quality and validity
- Data synthesis and analysis
- Results
- Discussion
- *Funding or sponsorship*

Randomized Clinical Trials

- Study question
- *Study population*
- *Randomization*
- *Blinding*
- *Interventions*
- *Outcomes*
- *Statistical analysis*
- Results
- Discussion
- *Funding or sponsorship*

(Key domains in *italics*.)

These latter two checklists excluded funding; the scale omitted data abstraction and had a "Partial" score for search strategy.

Randomized Clinical Trials

In evaluating systems concerned with RCTs, the authors reviewed 20 scales,[18, 24-42] 11 checklists,[12-14, 43-50] one component evaluation,[51] and seven guidancedocuments.[1, 11, 52-57] In addition, they reviewed 10 rating systems used by AHRQ's EPCs.[58-68]

The authors designated a set of high-performing scales or checklists pertaining to RCTs by assessing their coverage of the following seven domains (see Box A): study population, randomization, blinding, interventions, outcomes, statistical analysis, and funding or sponsorship. They concluded that eight systems for RCTs represent acceptable approaches that could be used today without major modifications.[14, 18, 24, 26, 36, 38, 40, 45]

Two systems fully addressed all seven domains[24, 45] and six addressed all but the funding domain.[14, 18, 26, 36, 38, 40] Two were rigorously developed,[38, 40] but the significance of this factor has yet to be tested.

Of the 10 EPC rating systems, most included randomization, blinding, and statistical analysis,[58-61,63-68] and five EPCs covered study population, interventions, outcomes, and results as well.[60,61,63,65,66]

Users wishing to adopt a system for rating the quality of RCTs will need to do so on the basis of the topic under study, whether a scale or checklist is desired, and apparent ease of use.

Observational Studies

Seventeen non-EPC systems concerned observational studies. Of these, the authors categorized four as scales[31,32,40,69] and eight as checklists.[12-14,45,47,49,50,70] They classified the remaining five as guidance documents.[1,71-74] Two EPCs used quality rating systems for evaluating observational studies; these systems were identical to those used for RCTs.

To arrive at a set of high-performing scales or checklists pertaining to observational studies, the authors considered the following five key domains: comparability of subjects, exposure or intervention, outcome measurement, statistical analysis, and funding or sponsorship. As before, they concluded that systems that cover these domains represent acceptable approaches for assessing the quality of observational studies.

Of the 12 scales and checklists the authors reviewed, all included comparability of subjects either fully or in part. Only one included funding or sponsorship and the other four domains the authors considered critical for

Box B. Important Domains and Elements for Systems to Rate Quality of Individual Articles

Observational Studies

- Study question
- Study population
- *Comparability of subjects*
- *Exposure or intervention*
- *Outcome measurement*
- *Statistical analysis*
- Results
- Discussion
- *Funding or sponsorship*

Diagnostic Test Studies

- *Study population*
- *Adequate description of test*
- *Appropriate reference standard*
- *Blinded comparison of test and reference*
- *Avoidance of verification bias*

(Key domains in *italics*.)

observational studies. Five systems fully included all four domains other than funding or sponsorship.[14, 32, 40, 47, 50]

Two EPCs evaluated observational studies using a modification of their RCT quality system.[60,64] Both addressed the empirically derived domain comparability of subjects, in addition to outcomes, statistical analysis, and results.

In choosing among the six high-performing scales for assessing study quality, users will have to evaluate which system is most appropriate for the task being undertaken, how long it takes to complete each instrument, and its ease of use. The authors were unable to evaluate these three instrument properties in the project.

Studies of Diagnostic Tests

Of the 15 non-EPC systems identified for assessing the quality of diagnostic studies, six are checklists.[12,14,49,75-78] Five domains are key for making judgments about the quality of diagnostic test reports: study population, adequate description of the test, appropriate reference standard, blinded comparison of test and reference, and avoidance of verification bias. Three checklists met all these criteria.[49,77,78] Two others did not address test description, but this omission is easily remedied should users wish to put these systems into practice.[12,14] The oldest system appears to be too incomplete for wide use.[75,76]

With one exception, the three EPCs that evaluated the quality of diagnostic test studies included all five domains either fully or in part.[59,68,79,80] The one EPC that omitted an adequate test description probably included this information apart from its quality rating measures.[79]

Systems for Grading the Strength of a Body of Evidence

The authors reviewed 40 systems that addressed grading the strength of a body of evidence: 34 from sources other than AHRQ EPCs and 6 from the EPCs. Their evaluation criteria involved three domains—quality, quantity, and consistency (Box C)—that are well-established variables for characterizing how confidently one can conclude that a body of knowledge provides information on which clinicians or policymakers can act.

The 34 non-EPC systems incorporated quality, quantity, and consistency to varying degrees. Seven systems fully addressed the quality, quantity, and consistency domains.[11,81-86] Nine others incorporated the three domains at least in part.[12,14,39,70,87-91]

Of the six EPC grading systems, only one incorporated quality, quantity, and consistency.[93] Four others included quality and quantity either fully or partially.[59, 60,67,68] The one remaining EPC system included quantity; study quality is measured as part of its literature review process, but this domain appears not to be directly incorporated into the grading system.[66]

Box C. Important Domains and Elements for Systems to Grade the Strength of Evidence

Quality: the aggregate of quality ratings for individual studies, predicated on the extent to which bias was minimized.
Quantity: magnitude of effect, numbers of studies, and sample size or power.
Consistency: for any given topic, the extent to which similar findings are reported using similar and different study designs

Discussion

Identification of Systems

The authors identified 1,602 articles, reports, and other materials from their literature searches, web searches, referrals from their technical expert advisory group, suggestions from independent peer reviewers of an earlier version of this report, and a previous project conducted by the RTI–UNC EPC. In the end, the authors' formal literature searches were the least productive source of systems for this report. Of the more than 120 systems they eventually reviewed that dealt with either quality of individual articles or strength of bodies of evidence, the searches *per se* generated a total of 30 systems that they could review, describe, and evaluate. Many articles from the searches related to study quality were essentially reports of primary studies or reviews that discussed "the quality of the data"; few addressed evaluating study quality itself.

The literature search was most problematic for identifying systems to grade the strength of a body of evidence. Medical Subject Headings (MeSH) terms were not very sensitive for identifying such systems or instruments. The authors attribute this phenomenon to the lag in development of MeSH terms specific for the evidence-based medicine field.

For those involved in evidence-based practice and research, the authors caution that they may not find it productive simply to search for quality rating or evidence grading schemes through standard (systematic) literature searches. This is one reason that the authors are comfortable with identifying a set of instruments or systems that meet reasonably rigorous standards for use in rating study quality and grading bodies of evidence. Little is to be gained by directing teams seeking to produce systematic reviews or technology assessments (or indeed clinical practice guidelines) to initiate wholly new literature searches in these areas.

At the moment, the authors cannot provide concrete suggestions for efficient search strategies on this topic. Some advances must await expanded options for coding the peer-reviewed literature. Meanwhile, the authors suggest that investigators wishing to build on these efforts might well consider tactics involving citation analysis and extensive contact with researchers and guideline developers to identify the rating systems they are presently using. In this regard, the efforts of at least some AHRQ-supported EPCs will be instructive.

Factors Important in Developing and Using Rating Systems

Distinctions Among Types of Studies, Evaluation Criteria, and Systems

The authors decided early on that comparing and contrasting study quality systems without differentiating among study types was likely to be less revealing or productive than assessing quality for systematic reviews, RCTs, observational studies, and studies of diagnostic tests independently. In the worst case, in fact, combining all such systems into a single evaluation framework risked nontrivial confusion and misleading conclusions, and they were not willing to take the chance that users of this report would conclude that "a single system" would suit all purposes. That is clearly not the case.

The authors defined quality based on certain critical domains, which comprised one or more elements. Some were based directly on empirical results that show that bias can arise when certain design elements are not met; they considered these factors as critical elements for the evaluation. Other domains or elements were based on best practices in the design and conduct of research studies. These are widely accepted methodologic standards, and investigators (especially for RCTs and observational studies) would probably be regarded as remiss if they did not observe them. The authors' evaluation of study quality systems was done, therefore, against rigorous criteria.

Finally, they contrasted systems on descriptive factors such as whether the system was a scale, checklist, or guidance document; how rigorously it was developed; whether instructions were provided for its use; and similar factors. This approach enabled the authors to home in on scales and checklists as the more likely methods for rating articles, that might be adopted more or less as is.

Numbers of Quality Rating Systems

The authors identified at least three times as many scales and checklists for rating the quality of RCTs as for other types of studies. Ongoing methodological work addressing the quality of observational and diagnostic test studies will likely affect both the number and the sophistication of these systems. Thus, the findings and conclusions with respect to these latter types of studies may need to be readdressed once results from more methodological studies in these areas are available.

Challenges of Rating Observational Studies

An observational study by its very nature "observes" what happens to individuals. Thus, to prevent selection bias, the comparison groups in an observation study are supposed to be as similar as possible except for the factors under study. For investigators to derive a valid result from their observational studies, they must achieve this comparability between study groups (and, for some types of prospective studies, maintain it by minimizing differential attrition). Because of the difficulty in ensuring adequate comparability between study groups in an observational study—both when the project is being designed or upon review after the work has been published—the authors raise the question of whether nonmethodologically trained researchers can identify when

potential selection bias or other biases more common with observational studies have occurred.

Instrument Length

Older systems for rating individual articles tended to be most inclusive for the quality domains the authors chose to assess.[24,45] However, these systems also tended to be very long and potentially cumbersome to complete. Shorter instruments have the obvious advantage of brevity, and some data suggest that they will provide sufficient information on study quality. Simply asking about three domains (randomization, blinding, and withdrawals) apparently can differentiate between higher- and lower-quality RCTs that evaluate drug efficacy.[34]

The movement from longer, more inclusive instruments to shorter ones is a pattern observed throughout the health services research world for at least 25 years, particularly in areas relating to the assessment of health status and health-related quality of life. Thus, this model is not surprising in the field of evidence-based practice and measurement. However, the lesson to be drawn from efforts to derive shorter, but equivalently reliable and valid, instruments from longer ones (with proven reliability and validity) is that substantial empirical work is needed to ensure that the shorter forms operate as intended. More generally, the authors are not convinced that shorter instruments *per se* will always be better, unless demonstrated in future empirical studies.

Reporting Guidelines

Reporting guidelines such as the CONSORT, QUOROM, and forthcoming STARD statements are not to be used for assessing the quality of RCTs, systematic reviews, or studies of diagnostic tests, respectively. However, the statements can be expected to lead to better reporting and two downstream benefits. First, the unavoidable tension (when assessing study quality) between the actual study design, conduct, and analysis and the reporting of these traits may diminish. Second, if researchers consider these guidelines at the outset of their work, they are likely to have better designed studies that will be easier to understand when the work is published.

Conflicting Findings When Bodies of Evidence Contain Different Types of Studies

A significant challenge arises in evaluating a body of knowledge comprising observational and RCT data. A contemporary case in point is the association between hormone replacement therapy (HRT) and cardiovascular risk. Several observational studies but only one large and two small RCTs have examined the association between HRT and secondary prevention of cardiovascular disease for older women with preexisting heart disease. In terms of quantity, the number of studies and participants is high for the observational studies and modest for the RCTs. Results are fairly consistent across the observational studies *and* across the RCTs, but between the two types of studies the results conflict. Observational studies show a treatment benefit, but the three RCTs showed no evidence that hormone therapy was beneficial for women with established cardiovascular disease.

Most experts would agree that RCTs minimize an important potential bias in observational studies, namely selection bias. However, experts also prefer more studies with larger aggregate samples and/or with samples that address more diverse patient populations and practice settings—often the hallmark of observational studies. The inherent tension between these factors is clear. The lesson the authors draw is that a system for grading the strength of evidence, in and of itself and no matter how good it is, may not completely resolve the tension. Users, practitioners, and policymakers may need to consider these issues in light of the broader clinical or policy questions they are trying to solve.

Selecting Systems for Use Today: A "Best Practices" Orientation

Overall, many systems covered most of the domains that are considered generally informative for assessing study quality. From this set, the authors identified 19 generic systems that fully address our key quality domains (with the exception of funding or sponsorship for several systems).[3,6-8,12,14,18,24,26,32,36,38,40,45,47,49,50,77,78] Three systems were used for both RCTs and observational studies.

In the authors' judgment, those who plan to incorporate study quality into a systematic review, evidence report, or technology assessment can use one or more of these 19 systems as a starting point, *being sure to take into account the types of study designs occurring in the articles under review.* Other considerations for selecting or developing study quality systems include the key methodological issues specific to the topic under study, the available time for completing the review (some systems seem rather complex to complete), and whether the preference is for a scale or a checklist. They caution that systems used to rate the quality of both RCTs and observational studies—what they refer to as "one size fits all" quality assessments—may prove to be difficult to use and, in the end, may measure study quality less precisely than desired.

The authors identified seven systems that fully addressed all three domains for grading the strength of a body of

evidence. The earliest system was published in 1994;[81] the remaining systems were published in 1999[11] and 2000,[82-86] indicating that this is a rapidly evolving field.

Systems for grading the strength of a body of evidence are much less uniform than those for rating study quality. This variability complicates the job of selecting one or more systems that might be put into use today. Two properties of these systems stand out. Consistency has only recently become an integral part of the systems reviewed in this area. The authors see this as a useful advance. Also continuing is the use of a study design hierarchy to define study quality as an element of grading overall strength of evidence. However, reliance on such a hierarchy without consideration of the domains discussed throughout this report is increasingly seen as unacceptable. As with the quality rating systems, selecting among the evidence grading systems will depend on the reason for measuring evidence strength, the type of studies that are being summarized, and the structure of the review panel. Some systems appear to be rather cumbersome to use and may require substantial staff, time, and financial resources.

Although several EPCs used methods that met the authors' criteria at least in part, these were topic-specific applications (or modifications) of generic parent instruments. The same is generally true of efforts to grade the overall strength of evidence. For users interested in systems deliberately focused on a specific clinical condition or technology, they refer readers to the citations given in the main report.

Recommendations for Future Research

Despite being able to identify various rating and grading systems that can more or less be taken off the shelf for use today, the authors found many areas in which information or empirical documentation was lacking. They recommend that future research be directed to the topics listed below, because until these research gaps are bridged, those wishing to produce authoritative systematic reviews or technology assessments will be somewhat hindered in this phase of their work. Specifically, they highlight the need for work on:

- Identifying and resolving quality rating issues pertaining to observational studies;
- Evaluating inter-rater reliability of both quality rating and strength-of-evidence grading systems;
- Comparing the quality ratings from different systems applied to articles on a single clinical or technology topic;
- Similarly, comparing strength-of-evidence grades from different systems applied to a single body of evidence on a given topic;
- Determining what factors truly make a difference in final quality scores for individual articles (and by extension a difference in how quality is judged for bodies of evidence as a whole);
- Testing shorter forms in terms of reliability, reproducibility, and validity;
- Testing applications of these approaches for "less traditional" bodies of evidence (i.e., beyond preventive services, diagnostic tests, and therapies)—for instance, for systematic reviews of disease risk factors, screening tests (as contrasted with tests also used for diagnosis), and counseling interventions;
- Assessing whether the study quality grids that the authors developed are useful for discriminating among studies of varying quality and, if so, refining and testing the systems further using typical instrument development techniques (including testing the study quality grids against the instruments they considered to be "high quality"); and
- Comparing and contrasting approaches to rating quality and grading evidence strength in the United States and abroad, because of the substantial attention being given to this work outside this country; such work would identify what advances are taking place in the international community and help determine where these are relevant to the U.S. scene.

Conclusion

The authors summarized more than 100 sources of information on systems for assessing study quality and strength of evidence for systematic reviews and technology assessments. After applying evaluative criteria based on key domains to these systems, they identified 19 study quality and seven strength of evidence grading systems that those conducting systematic reviews and technology assessment can use as starting points. In making this information available to the Congress and then disseminating it more widely, AHRQ can meet the congressional expectations set forth in the Healthcare Research and Quality Act of 1999 and outlined at the outset of the report. The broader agenda to be met is for those producing systematic reviews and technology assessments to apply these rating and grading schemes in ways that can be made transparent for groups developing clinical practice guidelines and other health-

related policy advice. The authors have also offered a rich agenda for future research in this area, noting that the Congress can enable pursuit of this body of research through AHRQ and its EPC program. They are confident that the work and recommendations contained in this report will move the evidence-based practice field ahead in ways that will bring benefit to the entire health care system and the people it serves.

References

1. Lohr KN, Carey TS. Assessing 'best evidence': issues in grading the quality of studies for systematic reviews. Joint Commission J Qual Improvement. 1999;25:470-479.
2. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999;282:1054-1060.
3. Barnes DE, Bero LA. Why review articles on the health effects of passive smoking reach different conclusions. JAMA.1998;279:1566-1570.
4. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol. 991;44:1271-1278.
5. Oxman AD, Guyatt GH, Singer J, et al. Agreement among reviewers of review articles. J Clin Epidemiol. 1991;44:91-98.
6. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med. 1994 Apr 15;120:667-676.
7. Sacks HS, Reitman D, Pagano D, Kupelnick B. Meta-analysis: an update. Mt Sinai J Med. 1996;63:216-224.
8. Auperin A, Pignon JP, Poynard T. Review article: critical review of meta-analyses of randomized clinical trials in hepatogastroenterology. Alimentary Pharmacol Ther. 1997;11:215-225.
9. Beck CT. Use of meta-analysis as a teaching strategy in nursing research courses. J Nurs Educ. 1997;36:87-90.
10. Smith AF. An analysis of review articles published in four anaesthesia journals. Can J Anaesth. 1997;44:405-409.
11. Clarke M., Oxman AD. Cochrane Reviewer's Handbook 4.0. The Cochrane Collaboration; 1999.
12. Khan KS, Ter Riet G, Glanville J, Sowden AJ, Kleijnen J. Undertaking Systematic Reviews of Research on Effectiveness. CRD's Guidance for Carrying Out or Commissioning Reviews: York, England: University of York, NHS Centre for Reviews and Dissemination; 2000.
13. New Zealand Guidelines Group. Tools for Guideline Development & Evaluation. Accessed July 10, 2000. Web Page. Available at: http://www.nzgg.org.nz/.
14. Harbour R, Miller J. A new system [Scottish Intercollegiate Guidelines Network (SIGN)] for grading recommendations in evidence based guidelines. BMJ. 2001;323:334-336.
15. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group. JAMA. 1994;272:1367-1371.
16. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. J Clin Epidemiol. 1995;48:167-171.
17. Cranney A, Tugwell P, Shea B, Wells G. Implications of OMERACT outcomes in arthritis and osteoporosis for Cochrane metaanalysis. J Rheumatol. 1997;24:1206-1207.
18. de Vet HCW, de Bie RA, van der Heijden GJMG, Verhagen AP, Sijpkes P, Kipschild PG. Systematic reviews on the basis of methodological criteria. Physiotherapy. June 1997;83(6):284-289.
19. Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. Lancet. 1998;351:47-52.
20. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. Health Technol Assess. 1998;2:1-276.
21. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet. 1999;354:1896-1900.
22. National Health and Medical Research Council (NHMRC). How to Use the Evidence: Assessment and Application of Scientific Evidence. Canberra, Australia: NHMRC; 2000.
23. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000;283:2008-2012.
24. Chalmers TC, Smith H Jr, Blackburn B, et al. A method for assessing the quality of a randomized control trial. Control Clin Trials. 1981;2:31-49.
25. Evans M, Pollock AV. A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. Br J Surg. 1985;72:256-260.
26. Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized control trials of primary treatment of breast cancer. J Clin Oncol. 1986;4:942-951.
27. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. Stat Med. 1989;8:441-454.
28. Gotzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. Control Clin Trials. 1989;10:31-56.
29. Kleijnen J, Knipschild P, ter Riet G. Clinical trials of homoeopathy. BMJ. 1991;302:316-323.
30. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol. 1992;45:255-265.

31. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. JAMA. 1994;272:101-104.
32. Goodman SN, Berlin J, Fletcher SW, Fletcher RH. Manuscript quality before and after peer review and editing at Annals of Internal Medicine. Ann Intern Med. 1994;121:11-21.
33. Fahey T, Hyde C, Milne R, Thorogood M. The type and quality of randomized controlled trials (RCTs) published in UK public health journals. J Public Health Med. 1995;17:469-474.
34. 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.
35. Khan KS, Daya S, Collins JA, Walter SD. Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. Fertil Steril. 1996;65:939-945.
36. van der Heijden GJ, van der Windt DA, Kleijnen J, Koes BW, Bouter LM. Steroid injections for shoulder disorders: a systematic review of randomized clinical trials. Brit J Gen Pract. 1996;46:309-316.
37. Bender JS, Halpern SH, Thangaroopan M, Jadad AR, Ohlsson A. Quality and retrieval of obstetrical anaesthesia randomized controlled trials. Can J Anaesth. 1997;44:14-18.
38. Sindhu F, Carpenter L, Seers K. Development of a tool to rate the quality assessment of randomized controlled trials using a Delphi technique. J Adv Nurs. 1997;25:1262-1268.
39. van Tulder MW, Koes BW, Bouter LM. Conservative treatment of acute and chronic nonspecific low back pain. A systematic review of randomized controlled trials of the most common interventions. Spine. 1997;22:2128-2156.
40. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health. 1998;52:377-384.
41. 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.
42. Turlik MA, Kushner D. Levels of evidence of articles in podiatric medical journals. J Am Podiatr Med Assoc. 2000;90:300-302.
43. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. N Engl J Med. 1982;306:1332-1337.
44. Poynard T, Naveau S, Chaput JC. Methodological quality of randomized clinical trials in treatment of portal hypertension. In Methodology and Reviews of Clinical Trials in Portal Hypertension. Excerpta Medica; 1987:306-311.
45. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. Pediatrics. 1989;84:815-827.
46. Imperiale TF, McCullough AJ. Do corticosteroids reduce mortality from alcoholic hepatitis? A meta-analysis of the randomized trials. Ann Intern Med. 1990;113:299-307.
47. Spitzer WO, Lawrence V, Dales R, et al. Links between passive smoking and disease: a best-evidence synthesis. A report of the Working Group on Passive Smoking. Clin Invest Med. 1990;13:17-42; discussion 43-46.
48. Verhagen AP, de Vet HC, 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;51:1235-1241.
49. National Health and Medical Research Council (NHMRC). How to Review the Evidence: Systematic Identification and Review of the Scientific Literature. Canberra, Australia : NHMRC; 2000.
50. Zaza S, Wright-De Aguero LK, Briss PA, et al. Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Task Force on Community Preventive Services. Am J Prev Med. 2000;18:44-74.
51. 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.
52. Prendiville W, Elbourne D, Chalmers I. The effects of routine oxytocic administration in the management of the third stage of labour: an overview of the evidence from controlled trials. Br J Obstet Gynaecol. 1988;95:3-16.
53. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA. 1994;271:59-63.
54. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1993;270:2598-2601.
55. The Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials. JAMA. 1994;272:1926-1931.
56. The Asilomar Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. Checklist of information for inclusion in reports of clinical trials. Ann Intern Med. 1996;124:741-743.
57. Moher D, Schulz KF, Altman DG, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. JAMA. 2001;285:1987-1991.
58. Aronson N, Seidenfeld J, Samson DJ, et al. Relative Effectiveness and Cost-Effectiveness of Methods of Androgen Suppression in the Treatment of Advanced Prostate Cancer. Evidence Report/Technology Assessment No. 4. Rockville, Md.: Agency for Health Care Policy and Research. AHCPR Publication No.99-E0012; 1999.
59. Lau J, Ioannidis J, Balk E, et al. Evaluating Technologies for Identifying Acute Cardiac Ischemia in Emergency Departments: Evidence Report/Technology Assessment: No. 26. Rockville, Md.: Agency for Healthcare Research and

Quality. AHRQ Publication No. 01-E006 (Contract 290-97-0019 to the New England Medical Center); 2000.

60. Chestnut RM, Carney N, Maynard H, Patterson P, Mann NC, Helfand M. Rehabilitation for Traumatic Brain Injury. Evidence Report/Technology Assessment No. 2. Rockville, Md.: Agency for Health Care Policy and Research. AHCPR Publication No. 99-E006; 1999.
61. Jadad AR, Boyle M, Cunningham C, Kim M, Schachar R. Treatment of Attention-Deficit/Hyperactivity Disorder. Evidence Report/Technology Assessment No. 11. Rockville, Md.: Agency for Healthcare Research and Quality. AHRQ Publication No. 00-E005; 1999.
62. Heidenreich PA, McDonald KM, Hastie T, et al. An Evaluation of Beta-Blockers, Calcium Antagonists, Nitrates, and Alternative Therapies for Stable Angina. Rockville, MD: Agency for Healthcare Research and Quality. AHRQ Publication No. 00-E003; 1999.
63. Mulrow CD, Williams JW, Trivedi M, Chiquette E, Aguilar C, Cornell JE. Treatment of Depression: Newer Pharmacotherapies. Evidence Report/Technology Assessment No. 7. Rockville, Md.: Agency for Health Care Policy and Research. AHRQ Publication No. 00-E003; 1999.
64. Vickrey BG, Shekelle P, Morton S, Clark K, Pathak M, Kamberg C. Prevention and Management of Urinary Tract Infections in Paralyzed Persons. Evidence Report/Technology Assessment No. 6. Rockville, Md.: Agency for Health Care Policy and Research. AHCPR Publication No. 99-E008; 1999.
65. West SL, Garbutt JC, Carey TS, et al. Pharmacotherapy for Alcohol Dependence. Evidence Report/Technology Assessment No. 5; Rockville, Md.: Agency for Health Care Policy and Research. AHCPR Publication No. 99-E004; 1999.
66. McNamara RL, Miller MR, Segal JB, et al. Management of New Onset Atrial Fibrillation. Evidence Report/Technology Assessement No.12. Rockville, Md.: Agency for Health Care Policy and Research; AHRQ Publication No. 01-E026; 2001.
67. Ross S, Eston R, Chopra S, French J. Management of Newly Diagnosed Patients With Epilepsy: A Systematic Review of the Literature. Evidence Report/Technology Assessment No. 39; Rockville, Md: Agency for Healthcare Research and Quality. AHRQ Publication No. 01-E-029; 2001.
68. Goudas L, Carr DB, Bloch R, et al. Management of Cancer Pain. Evidence Report/Technology Assessment. No. 35 (Contract 290-97-0019 to the New England Medical Center). Rockville, Md.: Agency for Health Care Policy and Research. AHCPR Publication No. 99-E004; 2000.
69. Corrao G, Bagnardi V, Zambon A, Arico S. Exploring the dose-response relationship between alcohol consumption and the risk of several alcohol-related conditions: a meta-analysis. Addiction. 1999;94:1551-1573.
70. Ariens GA, van Mechelen W, Bongers PM, Bouter LM, van der Wal G. Physical risk factors for neck pain. Scand J Work, Environ Health. 2000;26:7-19.
71. Carruthers SG, Larochelle P, Haynes RB, Petrasovits A, Schiffrin EL. Report of the Canadian Hypertension Society Consensus Conference: 1. Introduction. Can Med Assoc J. 1993;149:289-293.
72. Laupacis A, Wells G, Richardson WS, Tugwell P. Users' guides to the medical literature. V. How to use an article about prognosis. Evidence-Based Medicine Working Group. JAMA. 1994;272:234-237.
73. Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V. Users' guides to the medical literature. IV. How to use an article about harm. Evidence-Based Medicine Working Group. JAMA. 1994;271:1615-1619.
74. Angelillo IF, Villari P. Residential exposure to electromagnetic fields and childhood leukaemia: a meta-analysis. Bulletin of the World Health Organization. 1999;77:906-915.
75. Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. JAMA. 1984;252:2418-2422.
76. Arroll B, Schechter MT, Sheps SB. The assessment of diagnostic tests: a comparison of medical literature in 1982 and 1985. J Gen Intern Med. 1988;3:443-447.
77. Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests. Recommended Methods; 1996.
78. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-1066.
79. McCrory DC, Matchar DB, Bastian L, et al. Evaluation of Cervical Cytology. Rockville, Md.: Agency for Health Care Policy and Research. AHCPR Publication No.99-E010; 1999.
80. Ross SD, Allen IE, Harrison KJ, Kvasz M, Connelly J, Sheinhait IA. Systematic Review of the Literature Regarding the Diagnosis of Sleep Apnea. Rockville, Md.:Agency for Health Care Policy and Research; 1999.
81. Gyorkos TW, Tannenbaum TN, Abrahamowicz M, et al. An approach to the development of practice guidelines for community health interventions. Can J Public Health. Revue Canadienne De Sante Publique. 1994;85 Suppl 1:S8-13.
82. Briss PA, Zaza S, Pappaioanou M, et al. Developing an evidence-based Guide to Community Preventive Services---methods. The Task Force on Community Preventive Services. Am J Prev Med. 2000;18:35-43.
83. Greer N, Mosser G, Logan G, Halaas GW. A practical approach to evidence grading. Joint Commission J Qual Improv. 2000;26:700-712.
84. Guyatt GH, Haynes RB, Jaeschke RZ, et al. Users' Guides to the Medical Literature: XXV. Evidence-based medicine: principles for applying the Users' Guides to patient care. Evidence- Based Medicine Working Group. JAMA. 2000;284:1290-1296.
85. NHS Research and Development Centre of Evidence-Based Medicine. Levels of Evidence. Accessed January 12, 2001. Web Page. Available at: http://cebm.jr2.ox.ac.uk.
86. United States Surgeon General's Advisory Committee on Smoking and Health. Smoking and health: report of the advisory committee to the Surgeon General of the Public Health Service. Washington DC: U.S. Dept. of Health,

Education, and Welfare, Public Health Service, U.S Government Printing Office; 1964.

87. How to read clinical journals: IV. To determine etiology or causation. Can Med Assoc J. 1981;124:985-990.
88. Guyatt GH, Cook DJ, Sackett DL, Eckman M, Pauker S. Grades of recommendation for antithrombotic agents. Chest. 1998;114:441S-444S.
89. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. JAMA. 1995;274:1800-1804.
90. Hoogendoorn WE, van Poppel MN, Bongers PM, Koes BW, Bouter LM. Physical load during work and leisure time as risk factors for back pain. Scand J Work, Environ Health. 1999;25:387-403.
91. Sackett DL, Straus SE, Richardson WS, et al. Evidence-Based Medicine: How to Practice and Teach EBM. London: Churchill Livingstone; 2000.
92. Lohr KN. Grading Articles and Evidence: Issues and Options. Final Guidance Paper. Final report submitted to the Agency for Health Care Policy and Research for Contract No. 290-97-0011, Task 2. Research Triangle Park, N.C.: Research Triangle Institute; 1998.
93. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. Ann Intern Med. 1997;126:376-380.

Availability of Full Report

The full evidence report from which this summary was derived was prepared for the Agency for Healthcare Research and Quality by the Research Triangle Institute–University of North Carolina Evidence-based Practice Center under contract No. 290-97-0011. A limited number of prepublication copies of this report are available free of charge from the AHRQ Publications Clearinghouse by calling 800-358-9295. Requestors should ask for Evidence Report/Technology Assessment No. 47, *Systems to Rate the Strength of Scientific Evidence*. The final report is expected to be available by late Spring 2002 (AHRQ Publication No. 02-E016). At that time, printed copies may be obtained.

Internet users will be able to access the report online through AHRQ's Web site at: www.ahrq.gov/clinic/epcix.htm

www.ahrq.gov

AHRQ Pub. No. 02-E015
March 2002

ISSN 1530-440X