FACT SHEET

Rating the Strength of Scientific Research Findings

Agency for Healthcare Research and Quality • 2101 East Jefferson Street • Rockville, MD 20852

AHRQ is the lead Federal agency charged with supporting research designed to improve the quality of health care, reduce its cost, address patient safety and medical errors, and broaden access to essential services. AHRQ sponsors and conducts research that provides evidence-based information on health care outcomes; quality; and cost, use, and access. The information helps health care decisionmakers—patients and clinicians, health system leaders, and policymakers— make more informed decisions and improve the quality of health care services.

U.S. Department of Health and Human Services
Public Health Service

Introduction

Just as a jury needs evidence from reliable witnesses or forensic investigations to come to a just verdict, so reliable information is needed to help people answer questions about health care. For example, consumers inquire about which health plan to choose, clinicians inquire about which patients are most likely to benefit from specific treatments, health care organizations inquire about which services to provide and how, and policymakers inquire about whether and what kind of incentives can promote effective and safe health care. These and other health care decisions are increasingly being made with evidence from scientific research studies—evidence-based research—rather than relying on expert opinion or clinical experience alone. Thus, consumers, physicians, and other groups or organizations with a direct interest in health care issues need ready access to high-quality evidence that is clear and easy to understand.

New diagnostic and treatment options proliferate rapidly, and during the past decade the amount of health care information on the Internet and elsewhere has exploded. However, the presence of more information from scientific studies brings with it the challenge of developing effective approaches to identify *which* research evidence is of high quality and to navigate efficiently through the growing number of findings. This challenge is particularly difficult when different studies provide results that support different conclusions. Read in isolation, such studies only confuse those seeking to base their health care decisions on the best available evidence.

Research studies can shed light on whether a particular treatment works under controlled conditions (clinical research) or how well it performs in the health care system (health services research). Health services and clinical researchers have developed new methods to evaluate and synthesize the findings from multiple studies, based on systems that grade the quality of individual studies (which may be published in one or more scientific articles) or evaluate the strength of a body of evidence comprising many individual studies. These methods serve two purposes: to help evaluate the ever-increasing research literature and to enhance the literature's ability to be readily understood for decisionmaking across all sectors of health care.

Systematic reviews and technology assessments represent rigorous methods of compiling scientific evidence to

answer many health care questions, and they can help decisionmakers when similar studies present apparently conflicting results. Moreover, such assessments can help answer policy and other health care questions today, based on the current evidence, without the lengthy and expensive process of conducting additional large studies. For this reason, systematic reviews and technology assessments increasingly form the basis for individual and policy-level health care decisions. Systematic reviews and evidence-based technology assessments differ from traditional opinion-based narrative reviews in several ways. They attempt to minimize bias (systematic errors in the conduct of a research study) by the comprehensiveness and reproducibility of the search for and selection of articles under review. They also typically assess the methodologic quality of individual studies—that is, how well the studies were designed, conducted, and analyzed—and evaluate the overall strength of a body of evidence.

A Congressional Mandate

Through its Evidence-based Practice Center (EPC) program, which consists of EPCs in the United States and Canada, the Agency for Healthcare Research and Quality (AHRQ) advances our understanding of how to ensure that reviews of clinical or related health care literature are scientifically and clinically sound. Since 1999, the Agency has been mandated by Congress (in the Healthcare Research and Quality Act of 1999) to look at "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 was directed to make such methods or systems widely available.

To fulfill this charge from Congress, AHRQ commissioned the RTI International–University of North Carolina (RTI–UNC) EPC to undertake a study (*Systems to Rate the Strength of Scientific Evidence*, AHRQ Evidence-based Practice Report/Technology Assessment No. 47; AHRQ Publication No. 02-E016) that draws on its earlier work in this area. The new study also advances AHRQ's mission to improve the outcomes and quality of health care through research and dissemination of research results in the United States and other countries.

The EPC study has two main goals:

- Describe systems that rate the quality of evidence in individual studies, and in bodies of evidence that are an accumulation of studies addressing a common scientific issue.
- Provide guidance on current "best practices" in this field.

In their present study, the EPC researchers define quality as the extent to which a study's design, conduct, and analysis have minimized biases in selection of subjects ("selection bias") and measurement of outcomes ("measurement bias"), as well as differences in the study groups other than the factors being studied that might influence the results ("confounding bias"). Variable quality may affect analysts' or decisionmakers' confidence about findings from systematic reviews or technology assessments. Moreover, variable quality in efficacy or effectiveness studies may lead to conflicting results that complicate clinical and policy decisionmaking.

Methods

To carry out this study, the EPC researchers conducted two extensive literature searches and sought information from existing bibliographies, members of a technical

expert panel, and other sources to identify published research related to rating the quality of studies and the overall strength of evidence. They then developed and completed descriptive tables, which they called "grids," to compare and characterize existing systems. These grids focus on important categories that they concluded any acceptable rating instrument should cover. These elements of a rating system reflect steps in research design, conduct, or analysis that have been shown to protect against bias or other problems or that are long-accepted practices in epidemiology and related fields. When the researchers assessed the systems against these categories, they assigned scores of fully met (Yes), partially met (Partial), or not met (No).

Drawing on the results of their analysis, the researchers identified existing quality rating scales or checklists that systematic reviews and technology assessments can use, and they laid out the reasons for highlighting these specific instruments. Experts in the field and AHRQ staff provided an extensive peer review of the draft report, and the researchers incorporated these comments into the final publication.

Results

Data Collection

The researchers reviewed the titles and abstracts of 1,602 publications. From this set, they retained for this report 121 systems comprised of scales, checklists, other instruments, and guidance documents. Specifically, they assessed systems including 20 relating to systematic reviews, 49 systems for randomized controlled trials (RCTs), 19 for observational studies, and 18 for diagnostic test studies. Some systems applied to more than one type of study. In addition, they examined 40 systems for grading the strength of a body of evidence. For purpose of final evaluation, they focused on scales and checklists.

Systems for Rating the Quality of Individual Studies

Important Evaluation Categories and Elements

To evaluate systems related to rating the quality of individual studies, the authors defined important categories and elements for the four types of studies:

Systematic Reviews

Of 11 categories, the researchers designated seven as critical to adequately rate systematic reviews: *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 categories. A second checklist addressed all seven categories but merited only a "Partial" score for study question and study quality. Two other checklists addressed all categories but funding, and a scale omitted data abstraction and had a "Partial" score for search strategy.

Randomized Clinical Trials

The researchers designated scales or checklists as high-performing based on their coverage of seven critical categories: *study population, randomization, blinding, interventions, outcomes, statistical analysis,* and *funding or sponsorship*. The researchers concluded that eight systems for RCTs represent acceptable approaches that could be used today without major modifications. Two systems fully address all seven categories, and six address all but funding.

Most of the 10 EPC rating systems for randomized trials included three of the categories: *randomization, blinding,* and

Box A. Important Categories 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: number of studies, sample size or power, and magnitude of effect

Consistency: for any given topic, the extent to which similar findings are reported using similar and different study designs.

statistical analysis. Five EPCs also addressed *study population*, *interventions*, *outcomes*, and *results*.

Users wishing to adopt a system for rating the quality of RCTs will need to consider the topic under study, whether they prefer a scale or checklist, and ease of use of the system,

Observational Studies

Scales or checklists that represent acceptable approaches for assessing the quality of observational studies address five critical categories: *comparability of subjects*, *exposure or intervention*, *outcome measures*, *statistical analysis*, and *funding or sponsorship*. Of the 12 scales and checklists the researchers reviewed, all included comparability of subjects, either fully or in part. However, only one addressed all five critical categories. Five systems did not address funding or sponsorship but fully addressed the other four categories.

To choose among the six high-performing scales that assessed four or more categories, users should evaluate which system is most appropriate for their specific task, how long it takes to complete each instrument, and its ease of use. The researchers did not evaluate these factors in this study.

Diagnostic Test Studies

Five categories are critical to judge the quality of diagnostic test reports: *study population*, *adequate description of test*, *appropriate reference standard*, *blinded comparison of test and reference*, and *avoidance of verification bias*. Three checklists met all five categories. Two others did not address test description, but this omission is easily remedied. The oldest system seems too incomplete for wide use.

Systems for Grading the Strength of A Body of Evidence

The researchers reviewed 40 systems that addressed grading the strength of a body of evidence. Their evaluation criteria involved three categories—quality, quantity, and consistency (Box A). These criteria are well-established variables that characterize how reliably a body of knowledge provides information on which clinicians or policy makers can act.

The 40 systems incorporated quality, quantity, and consistency to varying degrees. Eight systems fully addressed the three domains. Nine others incorporated the three domains either fully or partially.

Identification of Systems

The researchers identified 1,602 articles, reports, and other materials from literature searches, Web searches, referrals from the 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. Formal literature searches generated only 30 of the 121 systems that they eventually reviewed. Many articles from the literature searches that related to study quality were essentially reports of primary studies or reviews to discuss "the quality of the data"; few of these articles addressed evaluating study quality itself.

The literature search had difficulty in identifying systems to grade the strength of a body of evidence. Medical Subject Headings (MeSH®) terms, used by MEDLINE®, were not very sensitive for identifying such systems or instruments. The researchers attribute this phenomenon to the lag in

development of MeSH terms specific to the evidence-based medicine field. Thus, they caution those involved in evidence-based practice and research that it may not be productive simply to use standard literature searches to find quality-rating or evidence-grading schemes. Teams producing systematic reviews or technology assessments (or, indeed, clinical practice guidelines) may not gain much by initiating new literature searches in these areas at this time.

Until options for coding the peer-reviewed literature are expanded, investigators wishing to build on the EPC researchers' efforts might well consider tactics such as citation analysis and extensive contact with researchers and guideline developers to identify the rating systems they use. In this regard, the researchers conclude, .the efforts of at least some EPCs will be instructive.

Factors Important in Developing And Using Rating Systems

Distinctions Among Types of Studies, Evaluation Criteria, and Systems

The researchers decided early on to differentiate among studies that assess systematic reviews, RCTs, observational studies, and diagnostic test studies. In the worst case, they felt that combining all such systems into a single evaluation framework risked significant confusion and misleading conclusions. Nor did they want users to assume—wrongly—that "a single system" suits all purposes.

The EPC researchers defined quality by the categories listed above, evaluating study quality systems against rigorous criteria. Some were based directly on empirical results showing that bias can arise when certain design elements are not met. They became the critical categories in each evaluation. Other categories or elements were based on best practices in the design and conduct of research studies. Investigators (especially for RCTs and observational studies) were expected to observe these widely accepted methodologic standards.

Finally, they compared systems on the basis of descriptive factors such as whether the system was a scale, a checklist, or a guidance document; how rigorously it was developed; and whether instructions were provided for its use. This approach led the researchers to conclude that scales and checklists are rating methods that users likely might adopt more or less as is.

Other Issues

The researchers examined a series of other issues, discussed in more detail in the report and the separately printed Technical Summary (AHRQ Publication No. 02-E015). Among these issues are the challenges of rating observational studies, the change over time in the length of rating instruments, and the role of reporting guidelines.

Observational studies present special problems for raters of quality. An observational study by its very nature "observes" what happens to individuals. 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. Without training in research methodology, investigators are likely to find it difficult to ensure adequate comparability between study groups in an observational study—both when the project is being designed and upon review after the work has been published.

Older systems for rating individual articles tended to be most inclusive for the categories of quality information the researchers assessed. However, these

systems are long and potentially cumbersome to complete. Shorter instruments have the obvious advantage of brevity, and some data suggest that they provide sufficient information on study quality. However, substantial empirical work is needed to ensure that the shorter forms operate as intended. The researchers report that they are not convinced that shorter instruments will always be better, unless future studies demonstrate such a finding.

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 diagnostic test studies, respectively. However, the guidelines can lead to better reporting and to two additional benefits. First, the unavoidable tension (when assessing study quality) between the actual study design, conduct, and analysis and the reporting about them may diminish. Second, if researchers consider these reporting guidelines as they begin their work, they may design studies that will be easier to understand when they publish their work.

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. The association between hormone (estrogen) replacement therapy (HRT) and cardiovascular risk illustrates this point. Several observational studies, but only one large and two small RCTs, have examined whether HRT can provide secondary prevention of heart disease in older women who already have such 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. However, results between the two types of studies conflict. Observational studies show a treatment benefit, but the three RCTs show no evidence that hormone therapy benefits women who had cardiovascular disease before treatment began.

Most experts agree that RCTs minimize selection bias, an important potential bias in observational studies. However, experts also prefer more studies with larger total numbers of subjects or with groups of subjects that address more diverse patient populations and practice settings—often the hallmark of observational studies. The inherent tension between these factors is clear. No system for grading the strength of evidence, no matter how good that system is, will completely resolve the tension. Users, practitioners, and policy makers may need to consider these issues in light of the broader clinical or policy questions they are trying to answer.

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

Many systems to rate the quality of individual studies covered the categories that the EPC researchers considered most important. The researchers identified 19 generic systems that fully address their key quality categories (with the exception of funding or sponsorship for several systems). Three systems were used for both RCTs and observational studies.

In their 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, *taking into account the types of study designs in the articles under review*. Users should base their choice on the topic under review, the available time to

complete the review (some systems seem rather complex to complete), and whether the users prefer a scale or checklist. The researchers 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 difficult to use and, in the end, may not measure study quality as precisely as desired.

To grade the strength of a body of evidence, the researchers identified eight systems that fully addressed all three categories. The earliest system was published in 1994; the rest were published in 1999 and 2000, indicating the rapid evolution of the field.

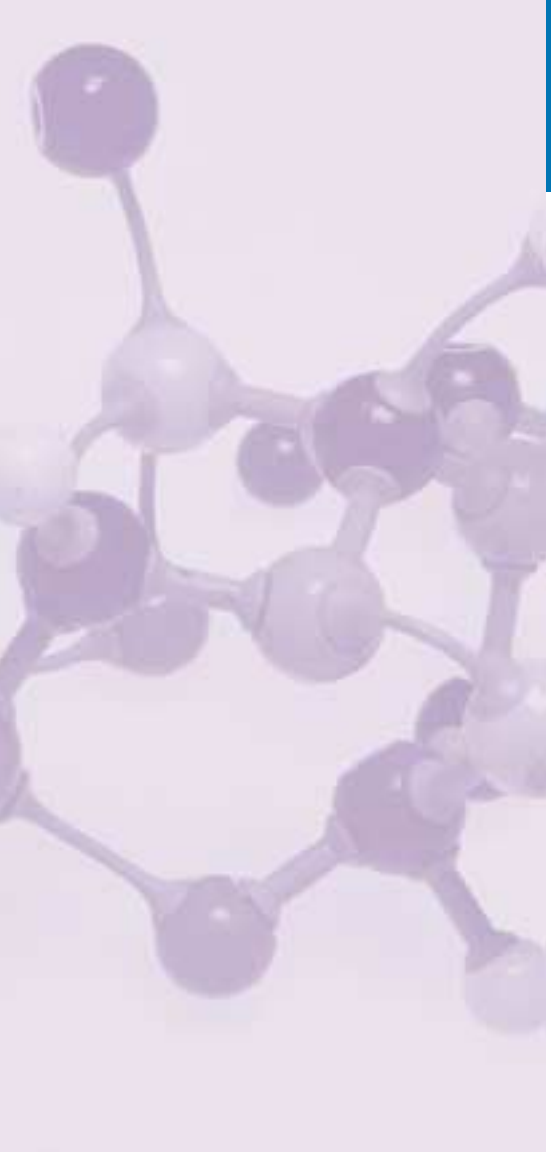
Systems for grading the strength of a body of evidence are much less uniform than those for rating study quality. This variability complicates the job of selecting one or more systems to use. Two properties of these systems stand out: Consistency has only recently become an integral part of the systems we reviewed, and this appears to be a useful advance. A study design hierarchy to define quality also persists as an element to grade overall strength of evidence. However, reliance on a hierarchy without consideration of the categories discussed throughout this report is not acceptable. As with the systems that rate individual studies, selecting among evidence grading systems will depend on the reason for measuring evidence strength, the type of studies being summarized, and the structure of the review panel. Some systems are cumbersome to use and may require substantial staff, time, and financial resources.

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

Recommendations for Future Research

Although various rating and grading systems can be used today, the researchers found many areas that lack information or empirical documentation. They recommend that future research address the topics below. Until these research gaps are bridged, authoritative systematic reviews or technology assessments will be somewhat limited. Specifically, the researchers highlight the need for work on the following areas:

- Identifying and resolving quality rating issues in observational studies;
- Evaluating reliability of both quality rating and strength-of-evidence grading systems when done by more than one rater;
- Comparing quality ratings from different systems applied to articles on a single clinical or technology topic;
- 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 studies (and, by extension, in how quality is judged for bodies of evidence as a whole);
- Testing shorter checklists or rating scales 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 were developed can discriminate among studies of varying quality and, if so, refining and testing the systems further (including testing the grids against the instruments that were rated "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 to this topic outside the United States, this work would identify advances in the international community that might be relevant to the U.S. scene.

Conclusion

The researchers summarized more than 100 sources of information on systems for assessing study quality and strength of evidence for systematic reviews and technology assessments. Using criteria based on key categories to these systems, they identified 19 study-quality and 7 strength-of-evidence grading systems that people conducting systematic reviews and technology assessment can use as starting points.

AHRQ not only sees this report as meeting the congressional mandate outlined earlier, but the Agency hopes that groups or organizations producing systematic reviews and technology assessments will apply these rating and grading schemes in a manner that will benefit groups developing clinical practice guidelines and other health-related policy advice. The report also offers a rich agenda for future research, which Congress may direct AHRQ and its EPC program to pursue. The work and recommendations in this report will undoubtedly move the field of evidence-based practice ahead in ways that will benefit the entire health care system and the people it serves.

Further Information

For additional information on the EPC report, *Systems to Rate the Strength of Scientific Evidence*, readers can request copies of the stand-alone Summary (AHRQ Publication No. 02-E015) or the full Report (AHRQ Publication No. 02-E016) from the AHRQ Publications Clearinghouse at 1-800-358-9295 or via E-mail at ahrqpubs@ahrq.gov. Further information about the EPC program or AHRQ's activities can be accessed online via the Agency's Web site at: www.ahrq.gov, where the text of the report and summary will be made available in electronic form.

AHRQ Pub. No. 02-P022
April 2002