Evidence Report/Technology Assessment

Number 47

Systems to Rate the Strength Of Scientific Evidence

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
2101 East Jefferson Street
Rockville, MD 20852
http://www.ahrq.gov

Contract No. 290-97-0011

Prepared by:
Research Triangle Institute–University of North Carolina
Evidence-based Practice Center
Research Triangle Park, North Carolina

Suzanne West, Ph.D., M.P.H.
Valerie King, M.D., M.P.H.
Timothy S. Carey, M.D., M.P.H.
Kathleen N. Lohr, M.D.
Nikki McKoy, B.S.
Sonya F. Sutton, B.S.P.H.
Linda Lux, M.P.A.

AHRQ Publication No. 02-E016
April 2002

Suggested Citation:
West S, King V, Carey TS, et al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute–University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality. April 2002.

Preface

The Agency for Healthcare Research and Quality (AHRQ), formerly the Agency for Health Care Policy and Research (AHCPR), through its Evidence-Based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To bring the broadest range of experts into the development of evidence reports and health technology assessments, AHRQ encourages the EPCs to form partnerships and enter into collaborations with other medical and research organizations. The EPCs work with these partner organizations to ensure that the evidence reports and technology assessments they produce will become building blocks for health care quality improvement projects throughout the Nation. The reports undergo peer review prior to their release.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality.

We welcome written comments on this evidence report. They may be sent to: Director, Center for Practice and Technology Assessment, Agency for Healthcare Research and Quality, 6010 Executive Blvd., Suite 300, Rockville, MD 20852.

John M. Eisenberg, M.D.
Director
Agency for Healthcare Research and Quality

Robert Graham, M.D.
Director, Center for Practice and Technology Assessment
Agency for Healthcare Research and Quality

The authors of this report are responsible for its content. Statements in the report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services of a particular drug, device, test, treatment, or other clinical service.

Acknowledgments

This study was supported by Contract 290-97-0011 from the Agency for Healthcare Research and Quality (AHRQ) (Task No. 7). We acknowledge the continuing support of Jacqueline Besteman, JD, MA, the AHRQ Task Order Officer for this project.

The investigators deeply appreciate the considerable support, commitment, and contributions from Research Triangle Institute staff Sheila White and Loraine Monroe.

In addition, we would like to extend our appreciation to the members of our Technical Expert Advisory Group (TEAG), who served as vital resources throughout our process. They are Lisa Bero, PhD, Co-Director of the San Francisco Cochrane Center, University of California at San Francisco, San Francisco, Calif.; Alan Garber, MD, PhD, Professor of Economics and Medicine, Stanford University, Palo Alto, Calif.; Steven Goodman, MD, MHS, PhD, Associate Professor, School of Medicine, Department of Oncology, Division of Biostatistics, Johns Hopkins University, Baltimore, Md.; Jeremy Grimshaw, MD, PhD, Health Services Research Unit, University of Aberdeen, Scotland; Alejandro Jadad, MD, DPhil, Director of the program in eHealth innovation, University Health Network, Faculty of Medicine, University of Toronto, Toronto, Canada; Joseph Lau, MD, Director, AHRQ Evidence-based Practice Center, New England Medical Center, Boston, Mass.; David Moher, MSc, Director, Thomas C. Chalmers Center for Systematic Reviews, Children's Hospital of Eastern Ontario Research Institute, Ontario, Canada; Cynthia Mulrow, MD, MSc, Founding Director of the San Antonio Evidence-based Practice Center, San Antonio, Texas, and Associate Editor, *Annals of Internal Medicine*; Andrew Oxman, MD, MSc, Director, Health Services Research Unit, National Institute of Public Health, Oslo, Norway; and Paul Shekelle, MD, MPH, PhD, Director, AHRQ Evidence-based Practice Center, RAND-Southern California, Santa Monica, Calif.

We owe our thanks as well to our external peer reviewers, who provided constructive feedback and insightful suggestions for improvement of our report. Peer reviewers were Alfred O. Berg, MD, MPH, Chairman, U.S. Preventive Services Task Force, and Professor and Chair, Department of Family Medicine, University of Washington, Seattle, Wash.; Deborah Shatin, PhD, Senior Researcher, United Health Group, Minnetonka, Minn.; Edward Perrin, PhD, University of Washington, Seattle, Wash.; Marie Michnich, DrPH, American College of Cardiology, Bethesda, Md.; Steven M. Teutsch, MD, MPH, Senior Director, Outcomes Research and Management, Merck & Co., Inc., West Point, Pa.; Thomas Croghan, MD, Eli Lilly, Indianapolis, Ind.; John W. Feightner, MD, MSc, FCFP, Chairman, Canadian Task Force on Preventive Health Care and St. Joseph's Health Centre for Health Care, London, Ontario, Canada; Steve Lascher, DVM, MPH, Clinical Epidemiologist and Research Manager in Scientific Policy and Education, American College of Physicians-American Society of Internal Medicine, Philadelphia, Pa.; Stephen H. Woolf, MD, MPH, Medical College of Virginia, Richmond, Va.; and Vincenza Snow, MD, Senior Medical Associate, American College of Physicians-American Society of Internal Medicine, Philadelphia, Pa. In addition, we would like to extend our thanks to the seven anonymous reviewers designated by AHRQ.

Finally, we are indebted as well to several senior members of the faculty at the University of North Carolina at Chapel Hill: Harry Guess, MD, PhD, of the Departments of Epidemiology and Statistics, and Vice President of Epidemiology at Merck Research Laboratories, Blue Bell, Pa.; Charles Poole, MPH, ScD, of the Department of Epidemiology; David Savitz, PhD, Chair, Department of Epidemiology; and Kenneth F. Schulz, PhD, MBA, School of Medicine and Vice President of Quantitative Methods, Family Health International, Research Triangle Park, N.C.

Structured Abstract

Objectives. Health care decisions are increasingly being made on research-based evidence, rather than on expert opinion or clinical experience alone. This report examines systematic approaches to assessing the strength of scientific evidence. Such systems allow evaluation of either individual articles or entire bodies of research on a particular subject, for use in making evidence-based health-care decisions. Identification of methods to assess health care research results is a task that Congress directed the Agency for Healthcare Research and Quality to undertake as part of the Healthcare Research and Quality Act of 1999.

Search Strategy. The authors built on an earlier project concerning evaluating evidence for systematic reviews. They expanded this work by conducting a MEDLINE search (covering the years 1995 to mid-2000) for relevant articles published in English on either rating the quality of individual research studies or on grading a body of scientific evidence. Information from other Evidence-based Practice Centers (EPCs) and other groups involved in evidence-based medicine (such as the Cochrane Collaboration Methods Group) was used to supplement these sources.

Selection of Studies. The initial MEDLINE search for systems for assessing study quality identified 704 articles, while the search on strength of evidence identified 679 papers. Each abstract was assessed by two reviewers to determine eligibility. An additional 219 publications were identified from other sources The first 100 Abstracts in each group were used to develop a coding system for categorizing the publications.

Data Collection and Analysis. From the 1,602 titles and abstracts reviewed for the report, 109 were retained for further analsis. In addition, the authors examined 12 reports from various AHRQ-supported EPCs. To account for differences in study designs—systematic reviews and meta-analyses, randomized controlled trials (RCTs), observational studies, and diagnostic studies—the authors developed four Study Quality Grids whose columns denote evaluations domains of interest, and whose rows are the individual systems, checklists, scales, or instruments. Taken together, the grids form “evidence tables” that document the characterisitics (strengths and weaknesses) of these different systems.

Main Results. The authors separately analyzed systems found in the literature and those in use by the EPCs. Five non-EPC checklists for use with systematic reviews or meta-analyses accounted for at least six of seven domains needed to be considered high-performing. For analysis of RCTs, the authors concluded that eight systems represented acceptable approaches that could be used without major modifications. Six high-performing systems were identified to evaluate observational studies. Five non-EPC checklists adequately dealt with studies of diagnostic tests. For assessment of the strength of a body of evidence, seven systems fully addressed the quality, quantity, and consistency of the evidence.

Conclusions. Overall, the authors identified 19 generic systems that fully address their key quality domains for a particular type of study. The authors also identified seven systems that address all three quality domains grading the strength of a body of evidence. The authors also recommended future research areas to bridge gaps where information or empirical documentation is needed. The authors hope that these systems will prove useful to those developing clinical practice guidelines or other health-related policy advice.

Contents

Summary 1

EVIDENCE REPORT

Chapter 1. Introduction 15
- Motivation for and Goals of the Present Study 15
- Systematic Reviews of Scientific Evidence 16
- Quality Assessments in Systematic Reviews 17
  - Defining Quality 19
  - Developing Quality Assessments 19
  - Concepts of Study Quality and Strength of Evidence 21
- Organization of This Report 25

Chapter 2. Methods 27
- Solicitation of Input and Data 27
- Literature Search 27
  - Preliminary Steps 27
  - Searches 28
  - Title and Abstract Review 28
- Development of Study Quality Grids 31
  - Number and Structure of Grids 31
  - Overview of Grid Development 31
  - Defining Domains and Elements for Study Quality Grids 35
  - Assessing and Describing Quality Rating Instruments 38
- Development of Evidence Strength Grid 41
  - Quality 41
  - Quantity 42
  - Consistency 43
- Abstraction of Data 44
- Preparation of Final Report 44

Chapter 3. Results 45
- Data Collection Efforts 45
  - Rating Study Quality 45
  - Grading the Strength of a Body of Evidence 46
- Findings for Systems to Rate the Quality of Individual Studies 46
  - Background 46
  - Rating Systems for Systematic Reviews 48
  - Rating Systems for Randomized Controlled Trials 51
  - Rating Systems for Observational Studies 58
  - Rating Systems for Diagnostic Studies 61
- Findings for Systems to Rate the Strength of a Body of Evidence 64
  - Background 64
  - Evaluation According to Domains and Elements 64
  - Evaluation of Systems According to Three Domains That Address the

Strength of the Evidence 65
Evaluation of Systems According to Domains Considered Informative for Assessing the Strength of a Body of Evidence 68

Chapter 4. Discussion 75
Data Collection Challenges 75
Conceptualization of the Project 76
Quality of Individual Articles 76
Strength of a Body of Evidence 78
Study Quality 79
Growth in Numbers of Systems 79
Development of Systems Appropriate for Observational Studies 79
Longer or Shorter Instruments 80
Reporting Guidelines 80
Strength of a Body of Evidence 81
Interaction Among Domains 81
Conflict Among Domains When Bodies of Evidence Contain Different Types of Studies 81
Systems Related or Not Related to Development of Clinical Practice Guidelines 82
Emerging Uses of Grading Systems 82
Limitations of the Research 83
Selecting Systems for Use Today: A "Best Practices" Orientation 84
Rating Article Quality 84
Rating Strength of Evidence 84
EPC Systems 85
Recommendations for Future Research 85
Summary and Conclusion 87

References 89

Appendixes 99
Appendix A: Approaches to Grading Quality and Rating Strength of Evidence Used by Evidence-based Practice Centers 101
Appendix B: Quality of Evidence Grids 113
Appendix C: Strength of Evidence Grids 131
Appendix D: Annotated Bibliography 163
Appendix E: Excluded Articles 177
Appendix F: Abstraction Forms 189
Appendix G: Glossary 195

Figures

Figure 1. Study Design Algorithm 20
Figure 2. Continuum From Study Quality Through Strength of Evidence to Guideline Development 24

Tables

Table 1. Key Distinctions Between Narrative and Systematic Reviews by Core Features of Such Reviews....18
Table 2. Systematic Search Strategy to Identify Instruments for Assessing Study Quality....29
Table 3. Systematic Search Strategy to Identify Systems for Grading the Strength of a Body of Evidence....30
Table 4. Coding System Applied at the Abstract Stage for Articles Indentified During the Focused Literature Search for Study Quality Grid and for Strength of Evidence Grid....30
Table 5. Study Contructs Believed to Affect Quality of Studies....32
Table 6. Items Used to Describe Instruments to Assess Study Quality....33
Table 7. Domains and Elements for Systematic Reviews....36
Table 8. Domains and Elements for Randomized Controlled Trials....37
Table 9. Domains and Elements for Observational Studies....39
Table 10. Domains and Elements for Diagnostic Studies....40
Table 11. Domains for Rating the Overall Strength of a Body of Evidence....42
Table 12. Number of Systems Reviewed for Four Types of Studies by Type of System, Instruments, or Document....46
Table 13. Evaluation of Scales and Checklists for Systematic Reviews by Specific Instrument and 11 Domains....49
Table 14. Evaluation of Scales and Checklists for Systematic Reviews by Instrument and Seven Key Domains....52
Table 15. Evaluation of Scales, Checklists, and Component Evaluations for Randomized Controlled Trials by Specific Instrument and 10 Domains....54
Table 16. Evaluation of Guidance Documents for Randomized Controlled Trials by Instrument and 10 Domains....55
Table 17. Evaluation of Scales and Checklists for Randomized Controlled Trials by Instrument and Seven Key Domains....57
Table 18. Evaluation of Scales and Checklists for Observational Studies by Specific Instrument and Nine Domains....59
Table 19. Evaluation of Scales and Checklists for Observational Studies by Instrument and Five Key Domains....62
Table 20. Evaluation of Scales and Checklists for Diagnostic Test Studies by Specific Instrument and Five Domains....63
Table 21. Extent to Which 34 Non-EPC Strength of Evidence Grading Systems Incorporated Three Domains of Quality, Quantity, and Consistency....66
Table 22. Number of Non-EPC Systems to Grade Strength of Evidence by Number of Domains Addressed, Primary Purpose for System Development, and Year of Publication....67
Table 23. Characteristics of Seven Systems to Grade Strength of Evidence....69

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."[1, p. 472] For purposes of this study, we hold quality to be the extent to which a study's design, conduct, and analysis have minimized selection, measurement, and confounding biases, with our assessment of study quality systems reflecting this definition.

We 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 our analysis of relevant systems and instruments identified through literature searches and other sources. We present a selected set of systems that we 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 we 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. We 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, we conducted two extensive literature searches and sought further information from existing bibliographies, members of a technical expert panel, and other sources. We then developed and completed descriptive tables—hereafter "grids"—that enabled us to compare and characterize existing systems. These grids focus on important domains and elements that we 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. We assessed systems against domains and assigned scores of fully met (Yes), partially met (Partial), or not met (No).

Then, drawing on the results of our analysis, we identified existing quality rating scales or checklists that in our 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 we revised that draft as part of the steps to produce this report.

Results

Data Collection

We reviewed the titles and abstracts for a total of 1,602 publications for this project. From this set, we 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, we reviewed 12 reports from various AHRQ-supported EPCs. In all, we considered 121 systems as the basis for this report.

Specifically, we 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, we focused on scales and checklists. In addition, we reviewed 40 systems that addressed grading the strength of a body of evidence (34 systems identified from our 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.

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 are in *Italics*)

Systems for Rating the Quality of Individual Articles

Important Evaluation Domains and Elements

For evaluating systems related to rating the quality of individual articles, we 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 we regarded as critical for a scale or checklist to cover before we 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, we 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, we 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. 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, we reviewed 20 scales,[18,24-42] 11 checklists,[12-14,43-50] one component evaluation,[51] and seven guidance documents.[1,11,52-57] In addition, we reviewed 10 rating systems used by AHRQ's EPCs.[58-68]

We 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. We 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.

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 are in *Italics*)

Observational Studies

Seventeen non-EPC systems concerned observational studies. Of these, we categorized four as scales[31,32,40,69] and eight as checklists.[12-14,45,47,49,50,70] We 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, we considered the following five key domains: comparability of subjects, exposure or intervention, outcome measurement, statistical analysis, and funding or sponsorship. As before, we concluded that systems that cover these domains represent acceptable approaches for assessing the quality of observational studies.

Of the 12 scales and checklists we reviewed, all included comparability of subjects either fully or in part. Only one included funding or sponsorship and the other four domains we considered critical for observational studies.[45] 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, one 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. We were unable to evaluate these three instrument properties in the project.

Studies of Diagnostic Tests

Of the 15 non-EPC systems we 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]

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

Systems for Grading the Strength of a Body of Evidence

We 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. Our evaluation criteria involved three domains—quality, quantity, and consistency (Box C)—that are well-established variables for characterizing how confidently we 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]

Discussion

Identification of Systems

We identified 1,602 articles, reports, and other materials from our literature searches, web searches, referrals from our 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, our formal literature searches were the least productive source of systems for this report. Of the more than 120 systems we 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 we 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.

Our 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. We 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, we 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 we 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, we cannot provide concrete suggestions for efficient search strategies on this topic. Some advances must await expanded options for coding the peer-reviewed literature. Meanwhile, investigators wishing to build on our 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

We 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 we 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.

We 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; we 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. They 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. Our evaluation of study quality systems was done, therefore, against rigorous criteria.

Finally, we 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 us 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

We 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, our 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—we 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 we 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, we 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 we 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 we considered generally informative for assessing study quality. From this set, we 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.[14,40,45]

In our 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. We caution that systems used to rate the quality of both RCTs and observational studies—what we 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.

We 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 we reviewed in this area. We 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 our 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, we refer readers to the citations given in the main report.

Recommendations for Future Research

Despite our being able to identify various rating and grading systems that can more or less be taken off the shelf for use today, we found many areas in which information or empirical documentation was lacking. We 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, we 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 we 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 we 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

We 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, we 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. We 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. We 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.

Evidence Report

Chapter 1. Introduction

Throughout the 1990s and into the 21st century, the Agency for Healthcare Research and Quality (AHRQ, previously the Agency for Health Care Policy and Research [AHCPR]) 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. One special program has involved creating and funding a group of 12 Evidence-based Practice Centers (EPCs) in North America that specialize in producing systematic reviews (evidence reports and technology assessments) of the world's scientific and clinical literature and in enhancing the methods by which such work is done in a rigorous, yet efficient, manner. This report documents work done in 2000-2001 as part of the latter element of the Agency's mission—namely, advancing the field's understanding of how best to ensure that systematic reviews are scientifically and clinically robust.

Motivation for and Goals of the Present Study

In 1998, the Research Triangle Institute-University of North Carolina Evidence-based Practice Center (RTI-UNC EPC) prepared a report at the Agency's request to identify issues involved in assessing the quality of the published evidence.[1,92] The aim then was to provide AHRQ with information that would help all 12 EPCs ensure that the strength of the knowledge base about a given EPC topic was properly and adequately reflected in their final evidence reports. Lohr and Carey (1999) focused on ways to assess the quality of individual studies in systematic reviews; they found that many checklists, scales, and other similar tools were available for rating the quality of studies and that these tools varied widely.[1] They also reported that many tools were based on expert opinion, not grounded in empirical research; few scales used rigorous scale development techniques.

AHRQ asked the RTI-UNC EPC to undertake the present study, which extends and builds on the earlier report, for two reasons. The primary reason relates to a mandate from the Congress of the United States as part of the Healthcare Research and Quality Act of 1999, which created the Agency for Healthcare Research and Quality (AHRQ). This Act reauthorized the former AHCPR and extended many of its programs in quality of care, evidence-based practice, and technology assessment. Section 911(a) of Part B, Title IX, requires AHRQ, in collaboration with experts from the public and private sectors, to 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." The second reason for the current work relates to AHRQ's mission to support research that will improve the outcomes and quality of health care through research and dissemination. AHRQ's mission is being realized in part through its EPC program, the focus of which is "to improve the quality, effectiveness, and appropriateness of clinical care by facilitating the translation of evidence-based research findings into clinical practice." Thus, the research described in this report supports AHRQ's mission by providing information that EPCs and others can use to enhance research methods in the process of translating knowledge into practice.

The overarching goal of this project was to describe systems to rate the strength of scientific evidence focusing on methods used to conduct systematic reviews. The two specific aims were to:

- Conduct a rigorous review of quality scales, quality checklists, and study design characteristics (components) for rating the quality of individual articles.

- Identify and review methodologies for grading the strength of a body of scientific evidence—that is, an accumulation of many individual articles that address a common scientific issue.

We addressed these specific aims by conducting two focused literature searches, one for each specific aim, to identify published research related to these two issues. We then developed and completed descriptive tables or matrices—hereafter referred to as "grids"—to compare and characterize existing systems for assessing the quality of individual articles and rating the strength of bodies of evidence. In these preliminary stages, we solicited the advice and assistance of international experts. The grids and accompanying discussion form the results of this project. Drawing on the results of our analysis, we identified existing quality rating scales or checklists that in our view can be used in the production of systematic evidence reviews and technology assessments, along with a discussion of the reasons for highlighting these specific instruments. The mission of AHRQ's EPC program is carried out through the development of evidence reports and technology assessments—which collectively can be termed systematic reviews (as they are often known in the evidence-based practice field). For many in the clinical and policymaking communities, the products, indeed the lexicon, of evidence-based practice are unfamiliar, and one particular distinction may often be missed. This is the difference between a *systematic* review and the more familiar and more common narrative review. The next section of this chapter explicates the contrast between systematic and narrative reviews, with the aim of clarifying the significant role that systems for rating study quality and grading strength of evidence play in contemporary scientific endeavors of this sort.

Systematic Reviews of Scientific Evidence

What is a systematic review? According to Cook and colleagues (1997),[93] a systematic review is a type of scientific investigation of the literature on a given topic in which the "subjects" are the articles being evaluated. Thus, before a research team conducts a systematic review, it develops a well-designed protocol that lists: (1) a focused study question, (2) a specific search strategy, including the databases to be searched, and how studies will be identified and selected for the review according to inclusion and exclusion criteria, (3) the types of data to be abstracted from each article, and (4) how the data will be synthesized, either as a text summary or as some type of quantitative aggregation or meta-analysis. These steps are taken to protect the work against various forms of unintended bias in the identification, selection, and use of published work in these reviews.

In contrast, what is a narrative review? A narrative review is similar to a systematic review but without all the safeguards to control against bias. Table 1 (adapted from Cook et al.[95]) depicts the differences between systematic and narrative reviews. The major difference between these two approaches to synthesizing the clinical or scientific literature is that a systematic review attempts to minimize bias by the comprehensiveness and reproducibility of the search for and selection of articles for review.

The biases that can occur in systematic reviews are similar to those that are possible in clinical studies. For example, good study design for randomized controlled trials (RCTs) requires that allocation to treatment or control be randomized with the investigator "masked" (or

"blinded") to the subsequently assigned treatment (allocation concealment). This helps to ensure comparability of study groups and minimizes selection bias. By extension, in systematic reviews, if the literature search is not broad enough or the reasons for inclusion and exclusion of articles are not clearly specified, selection bias can arise in the choice of articles that are reviewed. [94,95] Another important difference between narrative reviews and systematic reviews is that systematic reviews typically assess how well the study was designed, conducted, and analyzed. That is, systematic reviews provide a measure of quality for each study (sometimes regarded as each article or publication) in the review. When research teams assemble the literature for a systematic review, it is important that they place more emphasis on the results from studies of higher rather than lower quality; this is an additional analytic step that does not typically occur in the conduct of narrative reviews. In addition, compared with traditional reviews, systematic reviews more typically provide explicit grading of the strength of the body of evidence in question.

The importance of taking a direct and explicit approach to assessing the quality of articles and strength of evidence lies, in part, in the need to be able to take account of differences in study quality and the impact of those differences on inferences that can be drawn about the scientific evidence. Empirical evidence indicates that the combined result or effect measure of interest in a review may be biased if studies of varying quality are summarized together.[51]

Quality Assessments in Systematic Reviews

The concern about study quality first arose in the early 1980s with the publication of a landmark paper by Chalmers and colleagues [24] and another extensive work by Hemminki, who evaluated the quality of trials done in 1965 through 1975 that were used to support the licensing of drugs in Finland and Sweden.[96] Since that time, numerous studies have provided evidence that study quality is important when producing systematic reviews. [51,97]

Table 1. Key Distinctions Between Narrative and Systematic Reviews, by Core Features of Such Reviews

Core Feature	Narrative Review	Systematic Review
Study question	Often broad in scope.	Often a focused clinical question.
Data sources and search strategy	Which databases were searched and search strategy are not typically provided.	Comprehensive search of many databases as well as the so-called gray literature. Explicit search strategy provided.
Selection of articles for study	Not usually specified, potentially biased.	Criterion-based selection, uniformly applied.
Article review or appraisal	Variable, depending on who is conducting the review.	Rigorous critical appraisal, typically using a data extraction form.
Study quality	If assessed, may not use formal quality assessment.	Some assessment of quality is almost always included as part of the data extraction process.
Synthesis	Often a qualitative summary.	Quantitative summary (meta-analysis) if the data can be appropriately pooled; qualitative otherwise.
Inferences	Sometimes evidence-based.	Usually evidence-based.

Source: Adapted from Cook et al., 1997.[93]

Thus, as this report will document, many quality scales and quality checklists have been developed in the past two decades or so for these evaluative purposes. In addition, several studies have appeared showing the importance of certain study design attributes or components, including randomization and double-blinding in the conduct of RCTs. These points are elaborated below and in Chapters 2 and 3. At this juncture, we note that the type of research being addressed in systematic reviews plays a major role in the conduct of those reviews and thus in the creation of systems for grading the evidence. Because of the significance of study design in this work, we present in Figure 1 a study design flow chart or algorithm (modified from Zaza et al.[50]) that discriminates among the various types of research published in the medical literature—RCTs, cross-sectional studies, case-control studies, and other so-called observational investigations.

Defining Quality

Critical to this discussion is the definition of quality, which authors of quality ratings often do not specify. In the previous AHRQ project, Lohr and Carey defined "methodologic quality" 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"; "nonmethodologic quality" refers to "the extent to which the information from a study has significant clinical or policy relevance." (Ref. 1, p. 472.)

We focus in this report on methodologic quality—that is, the extent to which a study's design, conduct, and analysis has minimized selection, measurement, and confounding biases. Our definition of quality refers to the internal validity of a study, not its external validity or generalizability. Although not all experts in the evidence-based practice field would take this approach, we consider issues of generalizability more relevant for developing clinical practice guidelines than for producing rigorous systematic reviews *per se*, in that guideline development is a step that occurs after a body of evidence on a clinical topic has been assembled and the overall strength of that evidence assessed.[98]

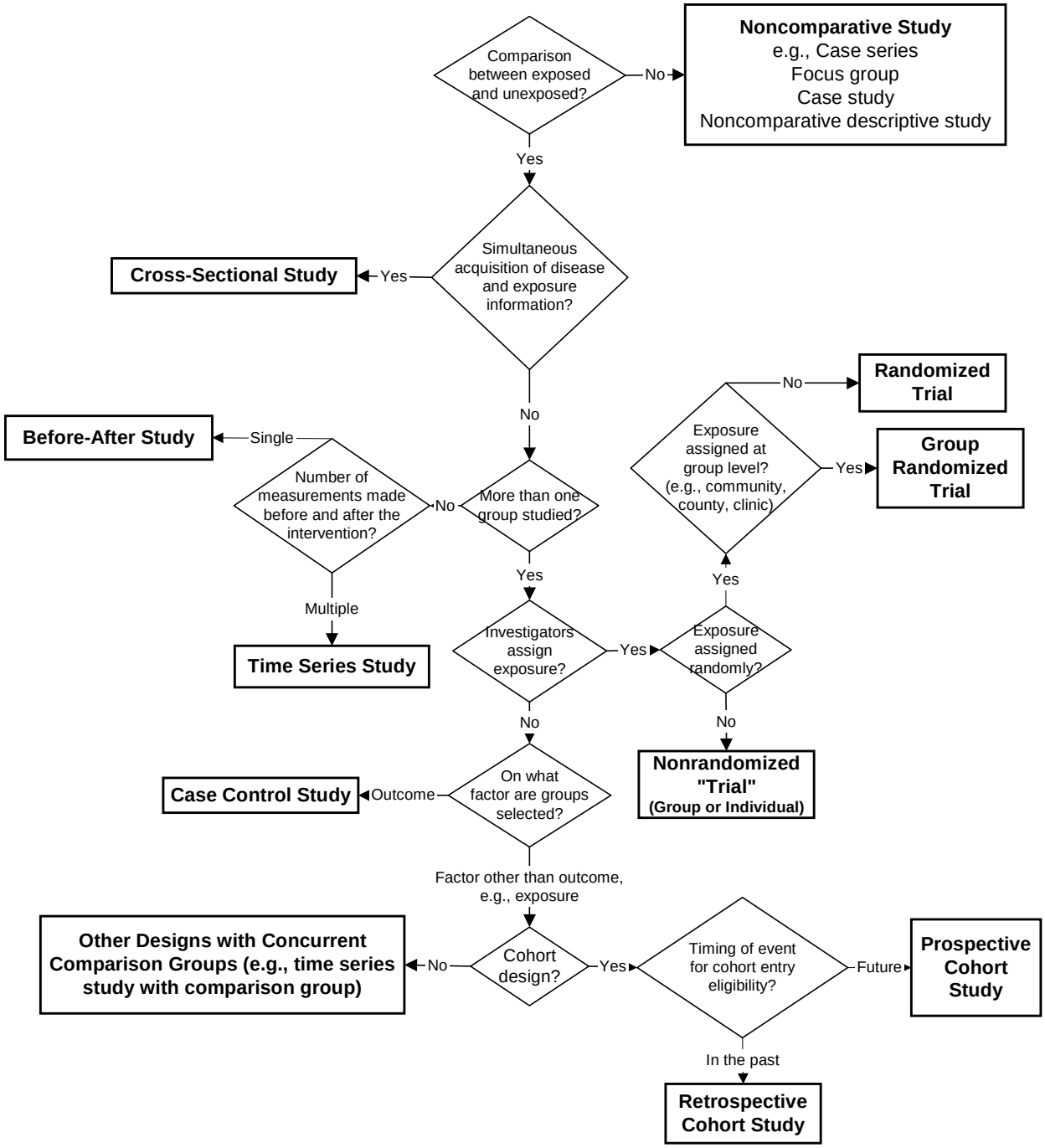
Developing Quality Assessments

Our project's first specific aim was to identify and compare tools to conduct quality assessment, which includes quality scales and checklists but also individual components of study quality. As part of the comparison among quality assessment instruments, it is important to understand proper scale development techniques.

Measurement scales, which in this context would include quality rating instruments, are developed in a stepwise fashion. The first steps involve defining the quality constructs or issues to be measured and the scope and purpose of the quality review, after which the actual questions are developed. The final steps require testing the instrument's reliability and validity and making such modifications as seem appropriate to meet conventional standards (for example, for internal consistency or test-retest reliability). [99]

Typically, instrument developers examine three types of validity: face, content, and criterion validity. Asking whether the instrument appears to measure what it was intended to measure assesses face validity. The extent to which the quality domain of interest—for example, randomization—is comprehensively assessed measures content validity. Criterion validity is

Figure 1. Study Design Algorithm



Modified from Zaza et al. 2000.[50]

defined as the extent to which the measurement correlates with an external criterion variable (a "gold standard"),[100] preferably one that can be measured objectively and independently. Khan and colleagues suggest that the last step in this iterative process is to determine the measurement properties of the quality instrument.[12] For many types of instruments, researchers measure criterion validity if an acceptable gold standard is available. Quality assessment tools of the type under consideration in this report have no true, "objective" gold standard that lies outside the domain of subjective assessment of the "goodness" of the study or article at hand. Because criterion validity cannot be assessed, some scale developers assess the instrument's reliability—that is, measuring whether a similar quality assessment score can be derived on the same study using either different scales or assessors (inter-rater reliability).[12] The rigorousness with which a scale is developed may influence its measurement properties.

Using Quality Ratings

No consensus exists on how study quality should be used in a systematic review. Moher, Jadad, and Tugwell (1996) describe four ways that quality assessment of RCTs may be used in systematic reviews and meta-analyses.[101] The most basic approach is to use quality as an inclusion threshold. Many reviewers, for example, admit only RCTs into a systematic review and eliminate other study designs from further consideration. Others have used a numeric quality score as a statistical weight when conducting meta-analyses (i.e., quantitative systematic reviews) to calculate a summary estimate of effect. A third method involves conducting a cumulative meta-analysis that is initiated by including only the higher quality studies and then adding studies of lesser quality sequentially. Finally, some have recommended that quality be examined visually in a plot.

Experts also disagree about whether quality should be formally scored, used as a threshold for inclusion or exclusion, employed in sensitivity analysis, applied in some other analytic framework, simply described, or not considered at all. Each approach has some potential advantages or some serious problems. If quality is to be used as a threshold for inclusion or exclusion, how quality is determined matters.[2] In systematic reviews of treatment, for instance, including a very poor quality study, regardless of its size, can profoundly influence summary estimates of the effects of that treatment. [41,97]

Complex statistical challenges arise when reviewers are attempting to arrive at a quantitative summary rather than attempting to conduct a more narrative review. Work by Detsky et al., Olkin, Moher et al., Sutton et al., and Tritchler is particularly helpful in guiding reviewers about the salient issues and statistical techniques involved. [20,30,101-103]

Concepts of Study Quality and Strength of Evidence

Study Quality

Systematic reviews comprise evidence based on research papers from the published literature and, whenever possible, from the "gray" or unpublished literature as well. Although much of the material identified and used for systematic reviews is from the peer-reviewed literature, the process and thoroughness of review conducted by journal reviewers and by those doing systematic reviews may not be the same. Thus, the literature compiled for systematic reviews

may be of varying quality, which can lead to conflicting summary estimates between systematic reviews on the same topic.

For example, Juni and colleagues evaluated study quality of 17 RCTs comparing low molecular weight heparin with standard heparin for preventing post-operative thrombosis using 25 different quality scales. Among the scales, both the indicators of study quality and their corresponding weights differed such that a study considered to be high quality on one scale was deemed low quality on another. Thus, summarizing the high quality articles according to one scale produced different relative risks than summarizing high quality studies using another scale.[2] Juni et al. found that an important predictor of summary relative risk was one particular component of study quality, whether the assessor of the outcome (risk of thrombosis) was masked to treatment allocation. This suggests that evaluating study quality is dependent on particular study design issues relevant to the topic under study. Their finding that a focus on methodologic components rather than summary scores to measure quality supports other work and editorial comment in the field.[104]

As discussed in more detail below, several published studies provide empirical evidence that inadequate description of certain elements of experimental study design—namely randomization procedures, allocation concealment (in which investigators do not know which drug will be assigned next), and outcome masking—have been associated with biased results.[2,51,105,106] Failure to mask the randomization procedures or outcome assessment was associated with elevated estimates of treatment effect compared with studies that reported using adequate masking procedures.

Whether potential design deficiencies in the published studies are the result of poor study design or poor reporting of study design is difficult to evaluate because reviewers typically see only the study report. Several collaborative efforts have put forth "statements" to standardize reporting; these include publishing guidelines for systematic reviews (QUOROM),[21] RCTs (CONSORT),[57] and observational studies (MOOSE).[23] These guidelines appear as checklists that authors can use to ensure that they have adequately addressed all the necessary components of a systematic review or publication on a given clinical or health services research project. Because of the evidence that poor quality studies may bias summary estimates from systematic reviews, researchers have developed and incorporated study quality assessment into their procedures for abstracting information from the literature and then describing and evaluating that literature. Numerous quality rating checklists and scales exist for RCTs. [101,107] Few instruments have been developed specifically for systematic reviews, observational studies, or investigations of diagnostic tests; however, most of those pertinent for observational studies of treatment effects are general enough to evaluate RCTs. Among existing instruments, even fewer scales and checklists have been developed using rigorous scale development techniques.

In this project, we compared and contrasted quality rating approaches using the definition of quality offered above, which is based on study design characteristics indicative of methodologic rigor. As explained in Chapter 2, Methods, we developed the grids for evaluating study quality using domains or specific items from various sources that described study quality or that discussed epidemiologic design standards. Some domains include explicit case definition specification, treatment allocation, control of confounding, extensiveness of follow-up, standardized and reproducible outcome assessment methods, and appropriate statistical analysis. Because design standards differ by study types (e.g., RCTs, observational studies, systematic reviews, and diagnostic studies), we developed one grid for each of these design types.

Strength of a Body of Evidence

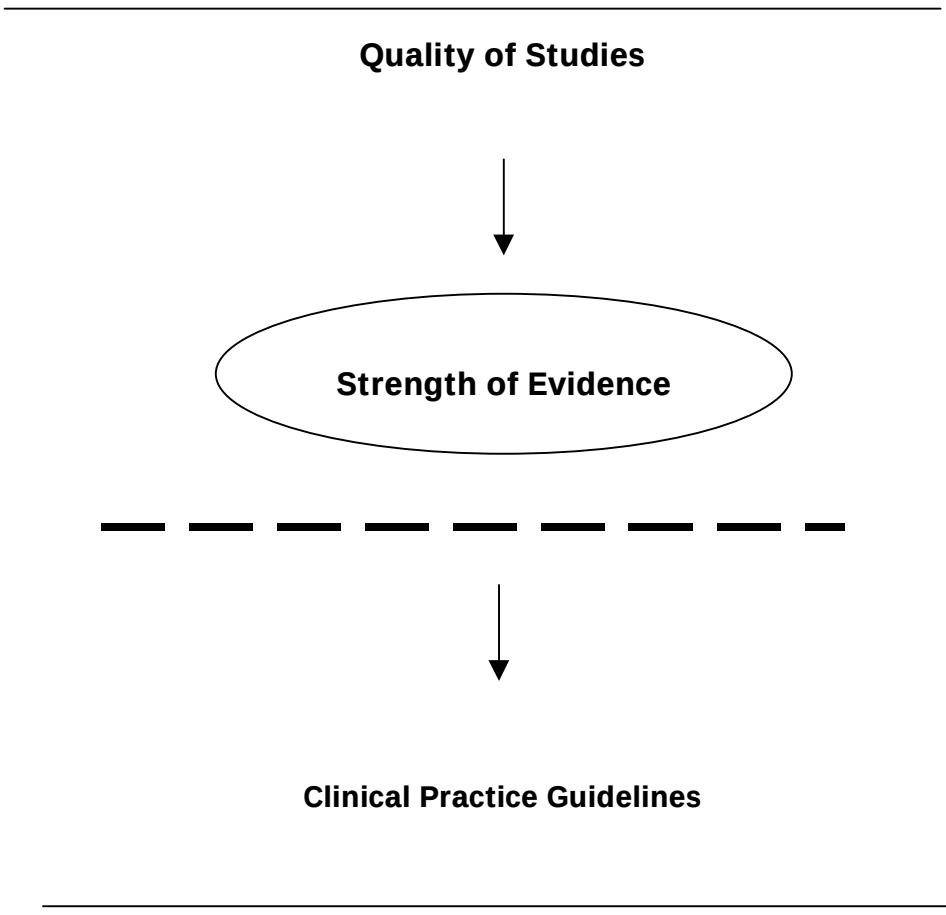
In conceptualizing this project, we contended that a continuum exists from rating study quality to grading the strength of a body of evidence. Grading the strength of a body of evidence incorporates judgments of study quality, but it also includes how confident one is that a finding is true and whether the same finding has been detected by others using different studies or different people. Thus, grading evidence strength stops at the dashed line in Figure 2. Only by incorporating population-specific information such as regional, racial, and clinical setting differences (akin to generalizability) does one derive a clinical or treatment guideline.

We extensively searched the literature to identify ways to grade the strength of a body of evidence. In the end, we determined that judging evidence strength does not typically appear to be a separate endeavor but rather is usually incorporated into the development of clinical practice guidelines and clinical recommendations within them. We thus limited our review of the guideline literature to the elements that address grading the strength of the evidence for a given topic *per se* and disregarded information addressing recommendation development.

In a manner analogous to the development of study quality grids, we created one additional matrix—an "evidence strength grid"—to capture the information concerning grading the strength of a body of scientific knowledge. In developing this grid, we posited that evaluating the strength of a body of evidence is similar to distinguishing between causal and noncausal associations in epidemiology.

Since the appearance of the Surgeon General's Report on Smoking and Health in 1964,[148,108] epidemiologists have been using five criteria for assessing causal relationships.[109] Two criteria, consistency and strength, are of particular relevance. C*onsistency* is the extent to which diverse approaches, such as different study designs or populations, for studying a relationship or link between a factor and an outcome will yield similar conclusions. S*trength* is the size of the estimated risk (of disease due to a factor) and its accompanying confidence intervals. Both of these concepts are directly related to grading the strength of a body of evidence.

Figure 2. Continuum from Study Quality Through Strength of Evidence to Guideline Development



The dashed line is the theoretical dividing line between summarizing the scientific literature and developing a clinical practice guideline. Below the dashed line, guideline developers would decide whether the evidence represents all the relevant subsets of the populations (or settings, or types of clinicians) for whom the guideline is being developed.

Other epidemiologic criteria such as *coherence*, which examines whether the cause-and-effect interpretation for an association conflicts with what is known of the natural history and biology of the disease, are more relevant for developing clinical recommendations. The remaining two causality criteria typically used in epidemiology, *specificity* and *temporality*, are more appropriate for measuring risk than for conducting technology assessments.

Based on these epidemiologic principles, the literature, and prior RTI-UNC EPC work, we concluded that grading the strength of a body of evidence should take three domains into account: quality, quantity, and consistency. *Quality* is defined as above, but in this case we are concerned with the quality of *all* relevant studies for a given topic. *Quantity* encompasses several aspects such as the number of studies that have evaluated the question, the overall sample size across all of the studies, and the magnitude of the treatment effect. Quantity is along the lines of "strength" from causality assessment and is typically reported in a comparative sense as a mean difference, relative risk, or odds ratio. *Consistency*—that is, whether investigations with both similar and different study designs report similar findings—can be assessed only if numerous studies are done. Thus, consistency is an important consideration when comparing one study with many individuals to several smaller studies with few individuals. We contend that one needs to address all three factors—quality, quantity, and consistency—when grading the strength of the evidence.

Organization of This Report

Chapter 2 of this report describes our technical approach, including methods for literature searches, interactions with outside experts and other EPCs, development of Study Quality and Evidence Strength Grids, and other steps. Appendix A describes our initial input from the EPCs. In Chapter 3 we present our results, including a detailed examination of the rating and grading systems we reviewed according to the domains that we regarded as significant for such systems to cover. Appendices B and C provide the actual completed grids by which to compare and contrast existing systems for assessing the quality of individual articles and grading the strength of bodies of evidence. Chapter 4 discusses our results in greater detail and provides a listing of several rating systems that, in our judgment, can be used for quality assessment purposes; it also offers our suggestions for future research. Appendix D gives an annotated bibliography of studies that provide empirical evidence on domains for rating study quality. The references include only studies cited in the body of this report; Appendix E cites excluded studies with the reason for exclusion. Appendix F contains an example of the electronic data abstraction tool we developed for this task. Appendix G provides a glossary of some of the terms we use in the context of this report.

Chapter 2. Methods

This project had numerous distinct tasks. We first solicited input and data from the Agency for Healthcare Research and Quality (AHRQ), its 12 Evidence-based Practice Centers (EPCs), and a group of international experts in this field. We then conducted an extensive literature search on relevant topics. From this information, we created tables to document important variables for rating and grading systems and matrices (hereafter denoted grids) to describe existing systems in terms of those variables. After analyzing and synthesizing all these data, we prepared this final report, which is intended to be appropriate for AHRQ to use in responding to the request from the Congress of the United States and in more broadly disseminating information about these systems and their uses in systematic reviews, evidence reports, and technology assessments.

As explained in Chapter 1, our ultimate goal was to create an integrated set of grids by which to describe and evaluate approaches and instruments for rating the quality of individual articles (referred to hereafter as Grids 1-4) and for grading the overall strength of a body of evidence (Grid 5). Here, we outline the project's overall methods, focusing on explicating the final set of grids. The completed grids can be found in Appendix B (Grids 1-4) and Appendix C (Grid 5).

Solicitation of Input and Data

Early in the project, we conducted a conference call with AHRQ to clarify outstanding questions about the project and to obtain additional background information. Enlisting the assistance and support of the other EPCs was a critical element of the effort. EPC directors or their designates participated in a second conference call in which we gave an overview of the project and discussed the information and documents we would need from them. We devised forms by which the EPCs could identify the methods they had used for rating the quality of the studies and grading the strength of the evidence in their AHRQ work or in similar activities for other sponsors (see Appendix A).

In addition, 10 experts served as a "technical expert advisory group" (TEAG; see Acknowledgments). We communicated with the TEAG through conference calls, occasional individual calls, and e-mail. Of particular importance were the TEAG members' efforts to clarify the conceptual model for the project, their identification of empirical work on study quality, and their review and critique of the grid structure. Eventually, several TEAG members also provided detailed reviews of a draft of this report.

Literature Search

Preliminary Steps

We carried out a multi-part effort to identify rating and grading systems and literature relevant to this question in several ways. First, we resurrected all documents acquired or generated in the original "grading project," including literature citations or other materials provided by the EPCs.[1] Second, as described in more detail below, we designed a supplemental literature search to identify articles that focused on generic instruments published in English

(chiefly from 1995 through mid-2000). Third, we used information from the EPC directors documenting the rating scales and classification systems that they have used in evidence reports or other projects for AHRQ or other sponsors. Fourth, we examined rating schemes or similar materials forwarded by TEAG members.

In addition, we tracked activities of several other groups engaged in examining these same questions. These include The Cochrane Collaboration Methods Group (especially work on assessing observational studies), the third (current) U.S. Preventive Services Task Force, and the Scottish Intercollegiate Guidelines Network (SIGN).

Finally, we reviewed the following international web sites for groups involved in evidence-based medicine or guideline development:

Canadian Task Force on Preventive Health Care (Canada), http://www.ctfphc.org/.
Centre for Evidence Based Medicine, Oxford University (U.K.), http://cebm.jr2.ox.ac.uk/ ;
National Coordination Centre for Health Technology Assessment (U.K.), http://www.ncchta.org/main.htm ;
National Health and Medical Research Council (Australia), http://www.nhmrc.health.gov.au/index.htm;
New Zealand Guidelines Group (New Zealand), http://www.nzgg.org.nz/; and
National Health Service (NHS) Centre for Reviews and Dissemination (U.K.), http://www.york.ac.uk/inst/crd/ ;
Scottish Intercollegiate Guidelines Network (SIGN) (U.K.), http://www.sign.ac.uk/ ;
The Cochrane Collaboration (international), http://www.cochrane.org/;

Searches

We searched the MEDLINE ® database for relevant articles published between 1995 and mid-2000 using the Medical Subject Heading (MeSH) terms shown in Tables 2 and 3 for Grids 1-4 (on rating the quality of individual studies) and Grid 5 (on grading a body of scientific evidence), respectively. For the Grid 5 search, we also had to use text words (indicated by an ".mp.") to make the search as inclusive as possible.

We compiled the results from all searches into a ProCite® bibliographic database, removing all duplicate records. We also used this bibliographic software to tag eligible articles and, for articles determined to be ineligible, to note the reason for their exclusion.

Title and Abstract Review

The initial search for articles on systems for assessing study quality (Grids 1-4) generated 704 articles (Table 2). The search on strength of evidence (Grid 5) identified 679 papers (Table 3).

We developed a coding system for categorizing these publications (Table 4) through two independent reviews of the abstracts from the first 100 articles from each search with consensus discussions as to whether each article should be included or excluded from full review. When abstracts were not available from the literature databases, we obtained them from the original article. The Project Director and the Scientific Director then independently evaluated the remaining titles and abstracts for the 604 articles (704 minus the 100 for coding system development) for Grids 1-4 and the 579 articles (679 minus the 100 for coding system

development) for Grid 5. Any disagreements were negotiated, erring on the side of inclusion as the most conservative approach.

We identified an additional 219 publications from various sources other than the formal searches, including the previous project,[1] bibliographies of seminal articles, suggestions from TEAG members, and searches of the web pages of groups working on similar issues (listed above). In all, we reviewed the abstracts for a total of 1,602 publications for the project; after review of all retained articles, we retained 109 that dealt with systems (i.e., scales, checklists, or other types of instruments or guidance documents) that were included in one or more of the grids and 12 EPC systems, for a total of 121 systems. The two-stage selection process that yielded these 121 systems is available from the authors on request.

Table 2. Systematic Search Strategy to Identify Instruments for Assessing Study Quality

	Search Strategy	Results
1	*Meta-analysis	895
2	*Randomized controlled trials/mt [Methods]	512
3	Systematic reviews.mp.	307
4	1 or 2 or 3	1,645
5	Limit 4 to (human and English language and year = 1995-2000)	858
6	Explode evidence-based medicine/ or explode quality control/ or explode reproducibility of results/ or explode data interpretation, statistical/ or explode "sensitivity and specificity"/ or explode research design/ or explode practice guidelines/	278,544
7	Explode guidelines/	13,969
8	Explode 8 (measurement scales or confidence profile or procedural methodology or study quality or study influence or effect measures).mp.	4,589
9	6 or 7 or 8	28,7281
10	5 and 9	704

* This term must be one of the four most important MeSH terms for the record.

Table 3. Systematic Search Strategy to Identify Systems for Grading the Strength of a Body of Evidence

	Search Strategy	Results
1	Explode evidence-based medicine/ or evidence.mp.	374,101
2	Strength or rigor or standards or authority or validity.mp.	111,824
3	1 and 2	7,323
4	*Randomized controlled trials/st [Standards]	308
5	3 or 4	7,621
6	Limit 5 to (human and English language and year = 1995-2000)	2,586
7	Grading.mp.	9,238
8	Explode observer variation/	8,697
9	Explode reproducibility of results/	52,017
10	Explode sensitivity and specificity/	87,492
11	Ranking.mp.	3,241
12	7 or 8 or 9 or 10 or 11	144,265
13	6 and 12	679†

* This term must be one of the four most important MeSH terms for the record.
† The figure of 679 articles identified excludes two publications that had been identified and counted for both the Study Quality and Evidence Strength literature searches.

Table 4. Coding System Applied at the Abstract Stage for Articles Identified During the Focused Literature Search for Study Quality Grid and for Strength of Evidence Grid

Codes Used for Both Study Quality Grids and Strength of Evidence Grids	Definition
Include	Obtain the full paper to assess whether it contains useful information for either grid
Ref-back	Reference or background paper to be obtained
Exclusions	
ECL	Editorial, comment, or letter
NR	Not relevant, requires specification of reason from below
NR-design/methods	Design or methodological issues, typically about clinical studies
NR-IA	Implementation/Application (e.g., described use of recommendations or guidelines in a clinical setting)
NR-OCD	Opinion/Commentary/Description (e.g., midway between ECL and review)
NR-ROS	Report of Study (e.g., report of a meta-analysis or clinical study)
NR-Review	Review/Overview (e.g., typically a narrative review of a clinical topic)
NR-Stat Meth	Statistical methodology (e.g., for conducting a meta-analysis)
NR-Other	Other reason for nonrelevance (e.g., continuing education, computer modeling systems)
Additional Code for Strength of Evidence Grid	
NR-Text word only (TWO)	Studies identified by text word but not relevant for inclusion (e.g., title or abstract had “evidence” or “recommend” as part of the text)

Development of Study Quality Grids
Number and Structure of Grids

We developed the four Study Quality Grids (Appendix B) to account for four different study designs—systematic reviews and meta-analyses, randomized controlled trials (RCTs), observational studies, and diagnostic studies.

Each Study Quality Grid has two parts. The first depicts the quality constructs and domains that each rated instrument covers; the other describes the instrument in various ways. For both Grids 1-4 (and Grid 5), columns denote evaluation domains of interest, and the rows are the individual systems, checklists, scales, or instruments. Taking these parts together, the grids form "evidence tables" that document the characteristics (strengths and weaknesses) of these different systems.

Overview of Grid Development

Preliminary Steps

Previous work done by the RTI-UNC EPC had identified constructs believed to affect the quality of studies (Table 5).[1] Beginning with these constructs and an annotated bibliography of scales and checklists for assessing the quality of RCTs,[101,107] we examined several of the more comprehensive systems of assessing study quality to settle on appropriate domains to use in the grids. These included approaches from groups such as the New Zealand Guidelines Group,[13] The Cochrane Collaboration,[11] the NHS Centre for Reviews and Dissemination,[85] and SIGN.[14] After three rounds of design, review, and testing, we settled on the domains and elements outlined in tables discussed below.

In addition to abstracting and assessing the content of quality rating instruments and systems, we gathered information on seven descriptive items for each article (Table 6). Definitions of key terms used in Table 6 appear in the glossary (Appendix G). These items, which were identical for all four study types, cover the following characteristics:

1. Whether the instrument was designed to be generic or specific to a given clinical topic;
2. The type of instrument (a scale, a checklist, or a guidance document);
3. Whether the instrument developers defined quality;
4. What method the instrument developers used to select items in the instrument;

Table 5. Study Constructs Believed to Affect Quality of Studies

Constructs	Definition
Selection of patients	• Who was included and who was excluded • Health, demographic, insurance, and other characteristics of these subjects • Diagnostic and/or prognostic criteria used
Comparability of study groups	• Randomization and allocation of patients to treatment and control/comparison groups • Similarity at baseline of these groups
Blinding	• Masking of patients, investigators, care providers, those who assessed outcomes to treatment groups or outcomes (or both)
Adequate sample size	• Size of the study • *A priori* justification of sample size • Consequent power
Therapeutic regimen	• Detailed information about the treatment, the settings in which the services were delivered, and the clinicians who delivered them • Description of co-interventions • Description of extra or unplanned treatments
Outcomes	• Choice of primary and secondary endpoints or outcomes • Ways the outcomes are measured
Availability of a study protocol	• Study administration, including length of follow-up period
Handling of withdrawals after eligibility determination	• Withdrawals, drop-outs, or other losses from the study, by patient group
Threats to validity	• Confounders and bias and how they are accounted for
Statistical analyses	• Appropriateness of statistical models • Adequacy of description and reporting of statistical analyses • Reporting levels of significance and/or confidence intervals • Extent to which all analyses that should have been done were done • "Intention-to-treat" analysis

Source: Adapted from Lohr and Carey (1999).[1]

Table 6. Items Used to Describe Instruments to Assess Study Quality

Descriptive Item*	Definitions of Descriptive Items
Generic or specific instrument	Generic: Instrument could be used to assess quality of any study of the type considered on that grid. Specific: Instrument is designed to be used to assess study quality for a particular type of outcome, intervention, exposure, test, etc.
Type of instrument	Scale: Instruments that contain several quality items that are scored numerically to provide a quantitative estimate of overall study quality. Checklist: Instruments that contain a number of quality items, none of which is scored numerically. Component: Individual aspect of study methodology (e.g., randomization, blinding, follow-up) that has a potential relation to bias in estimation of effect. Guidance Publication in which study quality is defined or Document: described, but does not provide an instrument that could be used for evaluative applications.
Quality concept discussion	Yes: Types or domains of quality that the instrument is designed to capture are discussed (e.g., biases that might affect the internal validity of the study). Partial: Quality concepts are discussed to some extent. No: Instrument itself or its documentation does not discuss the type or domains of study quality it assesses.
Method used to select items	Empiric: Items are based on criteria developed through empirical studies. Accepted: Items are based on accepted methodologic standards. Both: Items are of mixed empiric and accepted origin. Modification: The instrument represents a modification of another previously published instrument(s); original instrument is cited.
Rigor of development process	Yes: The use of standard scale development metrics in developing the instrument is explicitly described. Partial: The instrument was developed using an organized and reported consensus development process. No: No development process is reported or described.
Inter-rater reliability	Yes: Inter-rater reliability was assessed with appropriate statistical methods; results are reported in the grid. Partial: Issues concerning inter-rater reliability are discussed but the degree or range of reliability is not reported. No: Inter-rater reliability is not mentioned.
Instructions provided	Yes: Documentation of how to use and apply the instrument is adequate. Partial: Documentation of how to use the instrument is available in part (e.g., the questions on a checklist were clear and did not require substantial interpretation). No: Instrument did not provide instructions to guide its use.

* These items appear as column headings in the Study Quality and Evidence Strength Grids in Appendices B and C.

5. The rigor of the development process for this instrument;
6. Inter-rater reliability; and
7. Whether the developers had provided instructions for use of the instrument.

Domains and Elements for Evaluating Instruments to Rate Quality of Studies

A "domain" of study methodology or execution reflects factors to be considered in assessing the extent to which the study's results are reliable or valid (i.e., study quality). Each domain has specific "elements" that one might use in determining whether a particular instrument assessed that domain; in some cases, only one element defines a domain.
Tables 7-10 define domains and elements for the grids relevant to rating study quality. Although searching exhaustively for and cataloging evidence about key study design features and the risk of bias were steps beyond the scope of the present project, we present in Appendix D a reasonably comprehensive annotated bibliography of studies that relate methodology and study conduct to quality and risk of bias.

By definition, we considered all domains relevant for assessing study quality, but we made some distinctions among them. The majority of domains and their elements are based on generally accepted criteria—that is, they are based on standard "good practice" epidemiologic methods for that particular study design. Some domains have elements with a demonstrable basis in empirical research; these are designated in Tables 7-10 by italics, and we generally placed more weight on domains that had at least one empirically based element.

Empirical studies exploring the relationship between design features and risk of bias have often considered only certain types of studies (e.g., RCTs or systematic reviews), particular types of medical problems (e.g., pain or pregnancy), or particular types of treatments (e.g., antithrombotic therapy or acupuncture). Not infrequently, evidence from multiple studies of the "same" design factor (e.g., reviewer masking) comes to contradictory conclusions. Nevertheless, in the absence of definitive universal findings that can be applied to *all* study designs, medical problems, and interventions, we assumed that, when empirical evidence of bias exists for one particular medical problem or intervention, we should consider it in assessing study quality until further research evidence refutes it.

For example, we included a domain on funding and sponsorship of systematic reviews based on empirical work that indicates that studies conducted with affiliation to or sponsorship from the tobacco industry[3] or pharmaceutical manufacturers[110] may have substantial biases. We judged this to be sufficient evidence to designate this domain as empirically derived. However, we are cognizant that when investigators have strongly held positions, whether they be financially motivated or not, biased studies may be published and results of studies contrary to their positions may not be published. The key concepts are whether bias is likely to exist, how extensive such potential bias might be, and the likely effect of such bias on the results and conclusions of the study.

Although some domains have only a single element, others have several. To be able to determine whether a given instrument covered that domain, we identified elements that we considered "essential." Essential elements are those that a given instrument had to include before we would rate that instrument as having fully covered that domain. In Tables 7-10, these elements are presented in bold.

Finally, for domains with multiple elements, we specified the elements that the instrument had to consider before we would judge that the instrument had dealt adequately with that domain. This specification involved either specific elements or, in some cases, a count (a simple majority) of the elements.

Defining Domains and Elements For Study Quality Grids

Systematic Reviews and Meta-Analyses (Grid 1)

Table 7 defines the 11 quality domains and elements appropriate for systematic reviews and meta-analyses; these domains constitute the columns for Grid 1 in Appendix B. The domains are study question, search strategy, inclusion and exclusion criteria, interventions, outcomes, data extraction, study quality and validity, data synthesis and analysis, results, discussion, and funding or sponsorship. Search strategy, study quality and validity, data synthesis and analysis, and funding or sponsorship have at least one empirically based element. The remaining domains are generally accepted criteria used by most experts in the field, and they apply most directly to systematic reviews of RCTs.

Randomized Controlled Trials (Grid 2)

Table 8 presents the 10 quality domains for RCTs: study question, study population, randomization, blinding, interventions, outcomes, statistical analysis, results, discussion, and funding or sponsorship. Of these domains, four have one or more empirically supported elements: randomization, blinding, statistical analysis, and funding or sponsorship. Every domain has at least one essential element.

Observational Studies (Grid 3)

In observational studies, some factor other than randomization determines treatment assignment or exposure (see Figure 1 in Chapter 1 for clarification of the major types of observational studies). The two major types of observational studies are cohort and case-control studies. In a cohort study, a group is assembled and followed forward in time to evaluate an outcome of interest. The starting point for the follow-up may occur back in time (retrospective cohort) or at the present time (prospective cohort). In either situation, participants are followed to determine whether they develop the outcome of interest. Conversely, for a case-control study, the outcome itself is the basis for selection into the study. Previous interventions or exposures are then evaluated for possible association with the outcome of interest.

Table 7. Domains and Elements for Systematic Reviews

Domain	Elements*
Study Question	• **Question clearly specified and appropriate**
Search Strategy	• ***Sufficiently comprehensive and rigorous with attention to possible publication biases***
	• *Search restrictions justified (e.g., language or country of origin)*
	• Documentation of search terms and databases used
	• Sufficiently detailed to reproduce study
Inclusion and Exclusion Criteria	• **Selection methods specified and appropriate, with *a priori* criteria specified if possible**
Interventions	• **Intervention(s) clearly detailed for all study groups**
Outcomes	• **All potentially important harms and benefits considered**
Data Extraction†	• Rigor and consistency of process
	• Number and types of reviewers
	• Blinding of reviewers
	• Measure of agreement or reproducibility
	• Extraction of clearly defined interventions/exposures and outcomes for all relevant subjects and subgroups
Study Quality and Validity	• ***Assessment method specified and appropriate***
	• Method of incorporation specified and appropriate
Data Synthesis and Analysis	• ***Appropriate use of qualitative and/or quantitative synthesis, with consideration of the robustness of results and heterogeneity issues***
	• Presentation of key primary study elements sufficient for critical appraisal and replication
Results	• **Narrative summary and/or quantitative summary statistic and measure of precision, as appropriate**
Discussion	• **Conclusions supported by results with possible biases and limitations taken into consideration**
Funding or Sponsorship	• ***Type and sources of support for study***

* Elements appearing in italics are those with an empirical basis. Elements appearing in bold are those considered essential to give a system a Yes rating for the domain.
† Domain for which a Yes rating required that a majority of elements be considered.

Table 8. Domains and Elements for Randomized Controlled Trials

Domain	Elements*
Study Question	• **Clearly focused and appropriate question**
Study Population	• **Description of study population** • **Specific inclusion and exclusion criteria** • Sample size justification
Randomization	• *Adequate approach to sequence generation* • ***Adequate concealment method used*** • *Similarity of groups at baseline*
Blinding	• ***Double-blinding (e.g., of investigators, caregivers, subjects, assessors, and other key study personnel as appropriate) to treatment allocation***
Interventions	• **Intervention(s) clearly detailed for all study groups (e.g., dose, route, timing for drugs, and details sufficient for assessment and reproducibility for other types of interventions)** • Compliance with intervention • Equal treatment of groups except for intervention
Outcomes	• **Primary and secondary outcome measures specified** • Assessment method standard, valid, and reliable
Statistical Analysis	• ***Appropriate analytic techniques that address study withdrawals, loss to follow-up, missing data, and intention to treat*** • Power calculation • Assessment of confounding • Assessment of heterogeneity, if applicable
Results	• **Measure of effect for outcomes and appropriate measure of precision** • Proportion of eligible subjects recruited into study and followed up at each assessment
Discussion	• **Conclusions supported by results with possible biases and limitations taken into consideration**
Funding or Sponsorship	• ***Type and sources of support for study***

* Elements appearing in italics are those with an empirical basis. Elements appearing in bold are those considered essential to give a system a full Yes rating for the domain.

In all observational studies, selection of an appropriate comparison group of people without either the intervention/exposure or the outcome of interest is generally the most important and the most difficult design issue. Ensuring the comparability of the treatment groups in a study is what makes the RCT such a powerful research design. Observational studies are generally considered more liable to bias than RCTs, but certain questions can be answered only by using observational studies.

All nine domains and most of the elements for each domain apply generically to both cohort and case-control studies (Table 9). The domains are as follows: study question, study population, comparability of subjects, definition and measurement of the exposure or intervention, definition and measurement of outcomes, statistical analysis, results, discussion, and funding or sponsorship. Certain elements in the comparability-of-subjects domain are unique to case-control designs.

There are two empirically based elements for observational studies, use of concurrent controls and funding or sponsorship. However, a substantial body of accepted "best practices" exists with respect to design and conduct of observational studies, and we identified seven elements as essential.

Diagnostic Studies (Grid 4)

Assessment of diagnostic study quality is a topic of active current research.[78] We based the five domains in Table 10 for this grid on the work of the STARD (STAndards for Reporting Diagnostic Accuracy) group. The domains are study population, test description, appropriate reference standard, blinded comparison, and avoidance of verification bias. We designated five elements in Table 10 as essential, all of which are empirically derived.

The domains for diagnostic tests are designed to be used with the domains (and grids) for RCTs or observational studies because these are the basic study designs used to evaluate diagnostic tests. The domains for diagnostic tests can, in theory, also be applied to questions involving screening tests.

Assessing and Describing Quality Rating Instruments

Evaluating Systems According to Key Domains and Elements

To describe and evaluate systems for rating the quality of individual studies (Grids 1-4), we applied a tripartite evaluation scheme for the domains just described. Specifically, in the first part of each grid in Appendix B, we indicate with closed or partially closed circles whether the instrument fully or partially covered (respectively) the domain in question; an open circle denotes that the instrument did not deal with that domain. In the discussion that follows and in Chapter 3, we use the shorthand of "Yes," "Partial," and No" to convey these evaluations; in the grids they are shown as ●, ◐, ○, respectively.

Table 9. Domains and Elements for Observational Studies

Domains	Elements
Study Question	• **Clearly focused and appropriate question**
Study Population	• **Description of study populations** • Sample size justification
Comparability of Subjects†	<u>For all observational studies:</u> • Specific inclusion/exclusion criteria for all groups • Criteria applied equally to all groups • Comparability of groups at baseline with regard to disease status and prognostic factors • Study groups comparable to non-participants with regard to confounding factors • *Use of concurrent controls* • Comparability of follow-up among groups at each assessment <u>Additional criteria for case-control studies:</u> • Explicit case definition • Case ascertainment not influenced by exposure status • Controls similar to cases except without condition of interest and with equal opportunity for exposure
Exposure or Intervention	• **Clear definition of exposure** • Measurement method standard, valid and reliable • Exposure measured equally in all study groups
Outcome Measurement	• **Primary/secondary outcomes clearly defined** • Outcomes assessed blind to exposure or intervention status • Method of outcome assessment standard, valid and reliable • Length of follow-up adequate for question
Statistical Analysis	• Statistical tests appropriate • Multiple comparisons taken into consideration • Modeling and multivariate techniques appropriate • Power calculation provided • **Assessment of confounding** • Dose-response assessment, if appropriate
Results	• **Measure of effect for outcomes and appropriate measure of precision** • Adequacy of follow-up for each study group
Discussion	• **Conclusions supported by results with biases and limitations taken into consideration**
Funding or Sponsorship	• ***Type and sources of support for study***

* Elements appearing in italics are those with an empirical basis. Elements appearing in bold are those considered essential to give a system a Yes rating for the domain.

† Domain for which a Yes rating required that a majority of elements be considered.

Table 10. Domains and Elements for Diagnostic Studies

Domain	Elements*
Study Population	• ***Subjects similar to populations in which the test would be used and with a similar spectrum of disease***
Adequate Description of Test	• ***Details of test and its administration sufficient to allow for replication of study***
Appropriate Reference Standard	• ***Appropriate reference standard ("gold standard") used for comparison***
	• Reference standard reproducible
Blinded Comparison of Test and Reference	• Evaluation of test without knowledge of disease status, if possible
	• ***Independent, blind interpretation of test and reference***
Avoidance of Verification Bias	• ***Decision to perform reference standard not dependent on results of test under study***

* Elements appearing in italics are those with an empirical basis. Elements appearing in bold are those considered essential to give a system a Yes rating for the domain.

Yes evaluations mean that the instrument considered all or most of the elements for that domain and that it did not omit any element we defined as essential. A Partial rating meant that some elements in the domain were present but that at least one essential element was missing for that domain. No indicated that the instrument included few if any of the elements for a particular domain and that it did not assess any essential element.

Describing System Characteristics

Table 6 listed and defined the descriptive items that appear in the second part of each quality grid. We often had to infer certain pieces of information from the publications, as not all articles specified these descriptors directly. To say that a system had been "rigorously developed," we determined whether the authors indicated that they used typical instrument development techniques. We gave a Partial rating to systems that used some type of consensus panel approach for development.

Development of Evidence Strength Grid

The Strength of Evidence Grid (Grid 5, Appendix C) describes generic schemes for grading the strength of entire bodies of scientific knowledge—that is, more than one study evaluating the same or a similar relationship or clinical question about a health intervention or technology—rather than simply assessing the quality of individual articles. As discussed elsewhere, we have attempted to use criteria relevant to assessing a body of evidence without incorporating factors that are intended primarily to formulate, characterize, and support formal recommendations and clinical practice guidelines.

We defined three domains for rating the overall strength of evidence: quality, quantity, and consistency (Table 11). As with the Study Quality Grids, we have two versions. Grid 5A summarizes the more descriptive information from Grid 5B. In Grid 5A, we assigned a rating of Yes, Partial, or No (and applied the same symbols), depending on the extent to which the grading system incorporated elements of quality, quantity, and consistency.

Quality

Overall quality of a body of scientific studies is influenced by all the factors mentioned in our discussion of the quality of individual studies above. Grading systems that considered at least two of the following criteria—study design, conduct, analysis, or methodologic rigor—merited a Yes on quality. Systems that based their evidence grading on the hierarchy of research design without mention of methodologic rigor received a Partial rating.

Table 11. Domains for Rating the Overall Strength of a Body of Evidence

Domain	Definition
Quality	• The quality of all relevant studies for a given topic, where "quality" is defined as the extent to which a study's design, conduct, and analysis has minimized selection, measurement, and confounding biases
Quantity	• The magnitude of treatment effect • The number of studies that have evaluated the given topic • The overall sample size across all included studies
Consistency	• For any given topic, the extent to which similar findings are reported from work using similar and different study designs

Quantity

We use the construct "quantity" to refer to the extent to which there is a relationship between the technology (or exposure) being evaluated and outcome as well as to the amount of information supporting that relationship. Three main factors contribute to quantity:

- The magnitude of effect (i.e., estimated effects such as mean differences, odds ratio, relative risk, or other comparative measure);
- The number of studies performed on the topic in question (e.g., only a few versus perhaps a dozen or more); and
- The number of individuals studied, aggregated over all the relevant and comparable investigations, which provides the width of the confidence limits for the effect estimates.

The magnitude of effect is evaluated both within individual studies and across studies, with a larger effect indicative of a stronger relationship between the technology (or exposure) under consideration and the outcome. The finding that patients receiving a treatment are 5 times more likely to recover from an illness than those who do not receive the treatment is considered stronger evidence of efficacy than a finding that patients receiving a treatment are 1.3 times more likely to recover. However, absent any form of systematic bias or error in study design, and assuming equally narrow confidence intervals, there is no reason to consider this assertion (i.e., that the former is *stronger* evidence) to be the case. Rather, this illustrates the fact that one is simply measuring different sizes (magnitudes) of treatment effect. Nevertheless, no study is free from some element of potential unmeasured bias. The impact of such bias can overestimate or underestimate the treatment effect. Therefore, a large treatment effect partially protects an investigation against the threat that such bias will undermine the study's findings.

With respect to numbers of studies and individuals studied, common sense suggests that the greater the number of studies (assuming they are of good quality), the more confident analysts can be of the robustness of the body of evidence. Thus, we assume that systems for grading bodies of evidence ought to take account of the sheer size of that body of evidence.

Moreover, apart from the number of studies *per se* is the aggregate size of the samples included in those studies. All other things equal, a larger total number of patients studied can be expected to provide more solid evidence on the clinical or health technology question than a smaller number of patients. The line of reasoning is that hundreds (or thousands) of individuals included in numerous studies that evaluate the same issue give decisionmakers reason to believe

that that the topic has been thoroughly researched. In technical terms, the power of the studies to detect both statistically and clinically significant differences is enhanced when the size of the patient populations studied is larger.

However, a small improvement or difference between study patients and controls or comparisons must still be considered in light of the potential public health implications of the association under study. A minimal net benefit for study patients relative to comparison may seem insignificant except if it applies to very large numbers of individuals or can be projected to yield meaningful savings in health care costs. Thus, when using magnitude of an effect for judging the strength of a body of evidence, one must consider the size of the population that may be affected by the finding in addition to the effect size and whether it is statistically significant. Magnitude of effect interacts with number and aggregate size of the study groups to affect the confidence analysts can have in how well a health technology or procedure will perform. In technical terms, summary effect measures calculated from studies with many individuals will have narrower confidence limits than effect measures developed from smaller studies. Narrower confidence limits are desirable because they indicate that relatively little uncertainty attends the computed effect measure. In other words: a 95-percent confidence interval indicates that decisionmakers and clinicians can, with comfort, believe that 95 percent of the time the confidence interval will include (or cover) the true effect size.

A Yes for quantity meant that the system incorporated at least two of the three elements listed above. For example, if a system considered both the magnitude of effect and a measure of its precision (i.e., the width of the confidence intervals around that effect, which as noted is related to size of the studies), we assigned it a Yes. Rating systems that considered only one of these three elements merited a grade of Partial.

Consistency

Consistency is the degree to which a body of scientific evidence is in agreement with itself and with outside information. More specifically, a body of evidence is said to be consistent when numerous studies done in different populations using different study designs to measure the same relationship produce essentially similar or compatible results. This essentially means that the studies have produced reasonably reproducible results. In addition, consistency addresses whether a body of evidence agrees with externally available information about the natural history of disease in patient populations or about the performance of other or related health interventions and technologies. For example, information about older drugs can predict reactions to newer entities that have related chemical structures, and animal studies of a new drug can be used to predict similar outcomes in humans.

For evaluating schemes for grading strength of evidence, we treated the construct of consistency as a dichotomous variable. That is, we gave the instrument a Yes rating if it considered the concept of consistency and a No if it did not. No Partial score was given. Consistency is related to the concept of generalizability, but the two ideas differ in important ways. Generalizability (sometimes referred to as external validity) is the extent to which the results of studies conducted in particular populations or settings can be applied to different populations or settings. An intervention that is seen to work across varied populations and settings not only shows strong consistency but is likely to be generalizable as well. However, we chose to use consistency rather than generalizability in this work because we considered generalizability to be more pertinent to the further development of clinical practice guidelines (as

indicated in Figure 2, Chapter 1). That is, generalizability asks the question "Do the results of this study apply to my patient or my practice?" Thus, in assessing the strength of a body of literature, we de-emphasized the population perspective because of its link to guideline development and, instead, focused on the reproducibility of the results across studies.

Abstraction of Data

To abstract data on systems for grading articles or rating strength of evidence, we created an electronic data abstraction tool that could be used either in paper form (Appendix F) or as direct data entry. Two persons (Project Director, Scientific Director) independently reviewed all the quality rating studies, compared their abstractions, and adjudicated disagreements by discussion, additional review of disputed articles, and referral to another member of the study team as needed. For the strength of evidence work, the two principal reviewers each entered approximately half of the studies directly onto a template of the grid (Grid 5) and then checked each other's abstractions; again, disagreements were settled by discussion or additional review of the article(s) in question.

Preparation of Final Report

The authors of this report prepared two earlier versions. A partial "interim report" was submitted to AHRQ in the fall of 2000 for internal Agency use. More important, a draft final report was completed and submitted for wide external review early in 2001. A total of 22 experts and interested parties participated in this review; they included some members of the TEAG and additional experts invited by the RTI-UNC EPC team to serve in this capacity (see Acknowledgments) as well as several members of the AHRQ staff. This final report reflects substantive and editorial comments from this external peer review.

Chapter 3. Results

This chapter documents the results of this study in several parts. We first discuss the outcome of our data collection efforts (chiefly the two literature searches, one for rating study quality and the second for grading the strength of a body of evidence). We then provide our findings for rating study quality overall and by study type (i.e., systematic reviews, randomized controlled trials [RCTs], observational studies, and diagnostic studies). Last, we provide our findings on grading the strength of a body of evidence. Detailed tabular information is derived from the full assessments of all types of studies provided in Grids 1-4 (Appendix B) and Grid 5 (Appendix C); labels of domains of interest in developing the tables in this chapter are in some cases abbreviated versions of the domains defined in Tables 7-11 in Chapter 2 (e.g., funding or sponsorship is denoted funding).

For both study quality and strength of evidence, we identify selected systems that appear to cover domains we regard as particularly important. These systems might be regarded as ones that could be used today with confidence that they represent the current state of the art of assessing study quality or strength of evidence. Chapter 4, Discussion, examines the implications of these findings in more detail and gives our recommendations for research priorities concerned with systems for rating the scientific evidence for evidence reviews and technology assessments.

Data Collection Efforts

Rating Study Quality

Our first task was to identify instruments ("systems" in the original legislation mandating this report for the Agency for Healthcare Research and Quality [AHRQ]) for rating study quality. During our search process, we identified scales, checklists, and evaluations of quality components. In addition, we identified publications that discussed the importance of assessing article quality and that included quality items for consideration; we refer to these publications as guidance documents. To be complete, we include the guidance documents in Grids 1-4 (Appendix B), but in their current state we do not believe such documents can or should be used to rate the quality of individual studies.

Overall, we reviewed 82 different quality rating instruments or guidance documents for all four grids. This number encompasses reference papers that describe a study quality rating scheme or a rating method that is specific to work from an AHRQ-supported Evidence-based Practice Center (EPC). Because several of these 82 systems could be used to rate quality for more than one study design, we included them on multiple grids. Some came from our literature search, but we identified most by reviewing the previous effort of the Research Triangle Institute-University of North Carolina EPC[1] and work from Moher et al.[101] and by hand searching Internet sites and bibliographies.

As shown in Table 12, we assessed 20 systems for Grid 1 (systematic reviews), 49 systems for Grid 2 (RCTs), 19 for Grid 3 (observational studies), and 18 for Grid 4 (diagnostic studies). These systems can be characterized by instrument type as scales, checklists, or component evaluations; guidance documents; and EPC quality rating systems.

Grading the Strength of a Body of Evidence

We found it difficult to discern the most productive, yet specific, search terms for identifying literature that discussed grading a body of evidence. We approached our search from many different perspectives. In the end, although we identified numerous papers through the search, we found the majority of the relevant publications through hand searches and contacts with experts in the field. We suspect that, at present, the subject headings for coding the literature on this topic are not adequate to yield an appropriately thorough and productive search.

Thus, many of the 40 systems on which we provide information in Grid 5 (Appendix C) were identified through other sources or by reviewing bibliographies from papers retrieved by the search. Excluding the six evidence grading systems developed by the EPCs, approximately two-thirds (n = 22) of the remaining 34 systems arose from the guideline or clinical recommendations literature. Thus, only 12 of the evidence grading systems we reviewed were developed for *nonguideline* needs such as a literature synthesis or for purposes of evidence-based practice in general.

Findings for Systems to Rate The Quality of Individual Studies

Background

Chapter 2 describes the four study quality grids in Appendix B, including both the domains and elements used to compare rating systems (see Tables 7-10) and the properties used to describe them.

Evaluation According to Domains and Elements

The first part of each grid provides our assessment of the extent to which each system covered the relevant domains; we used a simple categorical scheme for this assessment:

- "Yes" (●, the system fully addressed the domain);
- "No" (○, it did not address the domain at all); or

Table 12. Number of Systems Reviewed for Four Types of Studies, by Type of System, Instrument, or Document

Study Design (Grid)	Total	Scales, Checklists, and Component Evaluations	Guidance Documents	EPC Rating Systems
Systematic Reviews (Grid 1)	20	11	9	0
Randomized Controlled Trials (Grid 2)	49	32	7	10
Observational Studies (Grid 3)	19	12	5	2
Diagnostic Tests (Grid 4)	18	6	9	3

- "Partial" (◐, it addressed the domain to some extent).

In defining domains, we differentiated between "empirical" elements and "good (or best) practice" elements. The former have been shown to affect the conduct and/or analysis of a study based on the results of rigorously designed methodological research. The latter elements have been identified as critical for the design of a well-conducted study but have not been tested in real life. As noted in Chapter 2 (and Appendix D), few empirical studies have been conducted; as a result, we have specified few empirical elements. Results of our analysis of each system appear below.

Description According to Key Characteristics

The second, descriptive part of each grid (see Table 6) provides general information on each rating system (e.g., type of system; whether inter-rater reliability had been assessed; how rigorously the system was developed). Although we focused on generic instruments, we did identify 18 "topic-specific" systems or instruments, especially among the EPC rating systems, and we also differentiate among the systems based on whether it is a scale, checklist, evaluation component only, or a guidance document.

Item Selection. In terms of approaches used by system developers to select the specific items or questions in their quality rating instruments, we found it difficult to determine whether they had chosen items on the basis of empirical research (theirs or others') or simply good practices (accepted) criteria. We based our categorization on whether the authors of the rating system referenced any empirical studies. One system included only empirical items;[34] another was a component evaluation of two empirical elements for RCTs (randomization and allocation concealment).[51] Remaining systems were based on accepted criteria, a mixture of accepted and empirical criteria, or modifications of another system.

Rigorous Development. As described in Chapter 1, a quality rating instrument could be developed in several steps, one of which is to measure inter-rater reliability. However, inter-rater reliability is only one facet of the instrument development process; by itself, it does not make an instrument "rigorously developed." We gave a system a Yes rating for rigorous development process if the authors indicated that they used "typical instrument development techniques," regardless of our rating for inter-rater reliability. Developmental rigor was typically a No for guidance documents, but we did give a Partial to some guidance documents because their quality criteria had been developed through formal expert consensus.

Inter-rater Reliability. Inter-rater reliability had been assessed in only 39 percent of the scales and checklists we reviewed, including those from the EPCs. We gave five systems (8 percent) a Partial rating for inter-rater reliability because the developers evaluated agreement among their raters but did not present the actual statistics. Inter-rater reliability was not relevant for guidance documents (always a No).

Quality Definition and Scoring. The last two descriptive items for quality rating systems—whether quality was defined or described and whether instructions were provided for use—had been included on an earlier summary of quality rating systems prepared by Moher and

colleagues.[107] Of the 82 systems we evaluated, 53 (65 percent) discussed their definition of quality to some extent (Yes or Partial for the category). Most of the systems did provide information on how to score each of the quality items; 64 systems (78 percent) were given either a Yes or Partial for instructions.

Rating Systems for Systematic Reviews

Type of System or Instrument

Twenty systems were concerned with systematic reviews or meta-analyses (Grid 1). Of these (Table 13), we categorized one as a scale[3] and 10 as checklists.[4-14] The remainder are considered guidance documents.[15-23,59,68] In the presentation below, we group scales and checklists into one set of results and comment on guidance documents separately.

Evaluation of Systems According to Coverage of Empirical or Best Practices Domains

Empirical Domains. The 11 domains used for assessing these systems (Table 13 or Grid 1A) reflect characteristics specific to both systematic reviews and general study design (see Table 7). Of these domains, four contain elements that are derived from empirical research: search strategy, study quality, data synthesis, and funding or sponsorship. Funding had only a single element (and it had an empirical basis). The study quality and data synthesis domains each comprised two or more elements (but only one element was empirically derived). Search strategy had four elements (of which two were empirical—comprehensive search strategy and justification of search restrictions). We give particular attention in the results below to the extent to which the systems we reviewed covered these empirical domains.

The one scale addressed all four domains with empiric elements (with a Partial grade for search strategy).[3] Of the 10 checklists, that by Sacks and colleagues fully addressed all four domains with empirical elements.[7] The checklist developed by Auperin and colleagues addressed three of the four empirically derived domains fully; the Partial score was for the study quality domain.[8]

All of the remaining eight systems excluded funding.[4-6,9-14] Five systems fully addressed three of the four empirically derived domains, omitting only funding.[4-6,11,12,14] The remaining three systems either did not address one or more empirically derived domains[9,13] or did so only partially.[10]

Best Practices Domains. The remaining seven domains—study question, inclusion and exclusion criteria, interventions, outcomes, data extraction, results, and discussion—come from best practices criteria. We included these for comparison purposes, mainly because many of the systems we evaluated included items addressing these domains.

The scale by Barnes and Bero fully addressed study question and inclusion/exclusion criteria but did not deal with or only partially addressed interventions, outcomes, data extraction, results, and discussion.[3] Of the 10 checklists, only one fully addressed all these good practices domains,[12] and two others addressed these domains to some degree.[7,8] The remaining seven systems entirely omitted one or more of these seven domains.[4-6,9-11,13,14]

Table 13. Evaluation of Scales and Checklists for Systematic Reviews, by Specific Instrument and 11 Domains

	Domains										
Instrument	**Study Question**	**Search Strategy***	**Inclusion/ Exclusion**	**Interventions**	**Outcomes**	**Data Extraction**	**Study Quality/ Validity***	**Data Synthesis and Analysis***	**Results**	**Discussion**	**Funding***
Oxman et al., 1991;[4] Oxman et al., 1991[5]	●	●	◐	○	◐	◐	●	●	●	○	○
Irwig et al., 1994[6]	●	●	●	●	●	●	●	●	●	○	○
Sacks et al., 1996[7]	●	●	●	●	●	●	●	●	●	◐	●
Auperin et al., 1997[8]	◐	●	●	●	●	●	◐	●	●	◐	●
Beck,1997[9]	●	●	●	○	○	●	○	●	●	●	○
Smith,1997[10]	◐	●	●	◐	○	○	●	◐	○	◐	○
Barnes and Bero, 1998[3]	●	◐	●	◐	○	○	●	●	◐	◐	●
Clarke and Oxman, 1999[11]	●	●	●	○	○	○	●	●	●	●	○
Khan et al., 2000[12]	●	●	●	●	●	●	●	●	●	●	○
New Zealand Guidelines Group, 2000[13]	●	●	●	○	●	○	○	○	●	◐	○
Harbour and Miller, 2001[14]	●	●	◐	●	●	○	●	●	●	○	○

*Domains with at least one element with an empirically demonstrated basis (see Table 7).

Every system addressed the inclusion/exclusion criteria at least partially. Most of these systems did cover study question and results, but the other domains excluded varied by system. One checklist did not address results in any way.[10] Four systems did not include intervention at all; [4,5,9,11,13] four did not include outcomes; [3,9-11] and five did not include data extraction.[3,10,11,13,14] The discussion domain was absent from four systems[4-6,14] and rated as Partial for five others.[3,7,8,10,13]

Because guidance documents have not been developed as tools for assessing quality *per se*, we did not contrast them with the scales and checklists and included them for illustrative purposes primarily. Like the scales and checklists, the results varied for the guidance documents. The two consensus statements that provide reporting guidelines include nearly all of the 11 domains. MOOSE included all 11 but received a Partial for the intervention domain.[23] The QUOROM statement did not include funding.[21]

Evaluation of Systems According to Descriptive Attributes

According to the descriptive information available in Grid 1B, none of the scales and checklists underwent rigorous development as defined earlier. We gave two checklists a score of Partial for this attribute,[11,14] mainly because the quality domains were selected by consensus. Four systems provided inter-rater reliability estimates that suggest that the quality ratings from multiple reviewers are consistent.[3-5,8,9] Interestingly enough, none of the systems that measured inter-rater reliability estimates had been rigorously developed.

Evaluation of Systems According to Seven Domains Considered Informative for Study Quality

Apart from the four domains that contained empirical elements, we concluded that three additional domains provide important information on the quality of a systematic review—study question, inclusion/exclusion criteria, and data extraction. The degree to which instruments concerned with systematic reviews covered these three domains is described just below, followed by a discussion of systems that appeared to deal with all seven domains.

Study Question. A clearly specified study question is important to define the search appropriately, determine which articles to exclude from the analysis, focus the interventions and outcomes, and conduct a meaningful data synthesis. Only two of the 20 systems omitted study question as a domain,[17,22] and an additional two received a Partial score for this domain.[8,10]

Inclusion/Exclusion. After the search is completed, determination of article eligibility is based on clearly specified selection criteria with reasons for inclusion and exclusion. Developing and adhering to strict inclusion and exclusion criteria makes the systematic review more reproducible and less subject to selection bias. Of the 20 systems we reviewed, every one addressed the inclusion/exclusion domain, with only three systems receiving a Partial for this domain.[4,5,14,15]

Data Extraction. How data had been extracted from single articles for purposes of systematic reviews is often overlooked in assessing the quality of a systematic review. Like the search strategy domain, the data extraction domain provides useful insight on the reproducibility of the systematic review. Reviews that do not use dual extraction may miss or misrepresent important

concepts. Of the 20 systems we reviewed, six omitted data extraction altogether [3,10,11,13,14,22] and three were given a Partial score for this domain.[4,5,15,19]

Coverage of Seven Key Domains. To arrive at a set of high-performing scales or checklists pertaining to systematic reviews, we took account of seven domains in all: study question, search strategy, inclusion/exclusion criteria, data extraction, study quality, data synthesis, and funding. We then used these seven domains as the criteria by which to identify a selected group of systems that could be said with some confidence to represent acceptable approaches that could be used today without major modifications. These are depicted in Table 14.

Five systems met most of the criteria for systematic reviews. One checklist fully addressed all seven domains.[7] A second checklist also addressed all seven domains but merited only a Partial for study question and study quality.[8] Two additional checklists[6,12] and the one scale[3] addressed six of the domains. These latter two checklists excluded funding; the scale omitted data extraction and had a Partial score for search strategy.

Rating Systems for Randomized Controlled Trials

Type of System or Instrument

In evaluating systems concerned with RCTs, we reviewed 20 scales,[18,24,42] 11 checklists,[12-14,43-50] one component evaluation,[51] and seven guidance documents.[1,11,52-57] In addition, we reviewed 10 EPC rating systems.[58-68] In the presentation below, we group scales, checklists, and the component system into a single set of results. We comment on guidance documents and EPC rating systems separately.

Our literature search focused on articles that described quality rating systems from 1995 until June 2000. Earlier work in this field had identified many scales and checklists for evaluating RCTs,[1,107] so duplicating prior work was not efficient. We did review and include many systems that we identified through the bibliographies of the more recent articles on RCT quality rating systems.

Evaluation of Systems According to Coverage of Empirical or Best Practices Domains

Empirical Domains. The 10 domains used for assessing these systems (Table 15 or Grid 2A) reflect characteristics specific to both RCTs and general study design (see Table 8). Of these domains, four contain elements that are derived from empirical research: randomization, blinding, statistical analysis, and funding or sponsorship. Both blinding and funding had only a single element (which was based on empirical research). The randomization domain comprised three elements, all of which were empirically derived. Statistical analysis had four elements, only one of which was empirically derived. In the results below, we focus on the extent to which the systems we reviewed covered these empirical domains.

Table 14. Evaluation of Scales and Checklists for Systematic Reviews by Instrument and Seven Key Domains

	Domains						
Instrument	**Study Question**	**Search Strategy***	**Inclusion/ Exclusion**	**Data Extraction**	**Study Quality***	**Data Synthesis/ Analysis***	**Funding***
Irwig et al., 1994[6]	●	●	●	●	●	●	○
Sacks et al., 1996[7]	●	●	●	●	●	●	●
Auperin et al., 1997[8]	◐	●	●	●	◐	●	●
Barnes and Bero,1998[3]	●	◐	●	○	●	●	●
Khan et al., 2000[12]	●	●	●	●	●	●	○

*Domains with at least one element with an empirically demonstrated basis (see Table 7).

Of the 32 scales, checklists, and component systems concerned with RCTs (Grid 2), only two fully addressed the four domains with empiric elements.[25,45] An additional 12 systems fully addressed randomization, blinding, and statistical analysis but not source of funding.[12,14,18,26,36,38-42,49,51] If we consider the systems that addressed the first three domains (randomization, blinding, statistical analysis) either partially or fully, we would add another 14 to this count.[13,25,27,28,29,31-35,37,43,44,47,48] Thus, only four of the RCT scales or checklists failed to address one or more of the three empirical domains, randomization, blinding, or statistical analysis.[29,30,46,50]

Best Practices Domains. The remaining six domains—study question, study population, interventions, outcomes, results, and discussion—come from best practices criteria. We included these for comparison purposes and because many of the systems we evaluated included items addressing these domains.

Focusing on the 14 scales, checklists, and component evaluation (Table 15) that fully addressed the three empiric domains—randomization, blinding, and statistical analysis—few systems included either study question or discussion.[14,38,40,45] However, 11 systems did address three other domains—study population, intervention, and results—either partially or fully.[12,14,18,24,26,36,38-40,42,45] Of these 11 systems, 10 also included outcomes as a domain; the exception is the work of the NHS Centre for Reviews and Dissemination.[12] Thus, these 11 systems included, either fully or in part, most of the domains that we selected to compare across systems.

Because guidance documents have not been developed as tools for assessing quality *per se*, we have examined them primarly for illustrative purposes (Table 16). The number of domains addressed in the guidance documents varied by system—from as few as three to all 10 of the domains. The consensus statements typically include most of the 10 domains.[55-57] The earliest consensus statement fully addressed seven domains, partially addressed one other, and failed to address two domains.[55] The Asimolar Working Group included all 10 domains but received a Partial for the randomization, blinding, and statistical analysis domains.[56] The most recent CONSORT statement fully addressed nine domains, omitting funding.[57]

Of the 10 EPC rating systems (see Grid 2A in Appendix B), all included both randomization and blinding at least partially. Statistical analysis was addressed either fully or partially by all but one system.[63] Study population, interventions, outcomes, and results were covered fully by five EPC systems.[60,61,63,65,66] EPC quality systems for RCTs rarely included either study question or discussion.

Evaluation of Systems According to Descriptive Attributes

The RCT system attributes are compared in Grid 2B (Appendix B). Most systems provided their definition of quality and selected their quality domains based on best practices criteria. Several used both best practices and empirical criteria for the selection process. Eight non-EPC scales and checklists were modifications of other systems.[26,27,31,33,35,37,41,44]

Table 15. Evaluation of Scales, Checklists, and Component Evaluations for Randomized Controlled Trials, by Specific Instrument and 10 Domains

	Domains									
Instrument	**Study Question**	**Study Population**	**Randomization***		**Interventions**	**Outcomes**	**Statistical Analysis***	**Results**	**Discussion**	**Funding***
Chalmers et al., 1981[24]	○	●	●	●	●	●	●	●	○	●
Liberati et al., 1986[26]	○	●	●	●	●	●	●	●	○	○
Reisch et al., 1989[45]	●	●	●	●	●	●	●	●	●	●
Schulz et al., 1995[51]	○	○	●	●	○	○	●	○	○	○
van der Heijden et al., 1996[36]	○	●	●	●	●	●	●	●	○	○
de Vet et al., 1997[18]	○	●	●	●	●	●	●	●	○	○
Sindhu et al., 1997[38]	●	●	●	●	●	●	●	◐	●	○
van Tulder et al., 1997[39]	○	◐	●	●	●	◐	●	●	○	○
Downs et al., 1998[40]	●	●	●	●	●	●	●	●	○	○
Moher et al., 1998[41]	○	○	●	●	○	○	●	○	○	○
Khan et al., 2000[12]	○	●	●	●	●	○	●	●	○	○
NHMRC, 2000[49]	○	○	●	●	○	●	●	○	○	○
Harbour and Miller, 2001[14]	●	●	●	●	●	●	●	●	○	○
Turlik et al., 2000[42]	○	●	●	●	◐	◐	●	●	○	○

*Domains with at least one element with an empirically demonstrated basis (see Table 8).

Table 16. Evaluation of Guidance Documents for Randomized Controlled Trials, by Instrument and 10 Domains

	Domains									
Instrument	Study Question	Study Population	Random-Ization*	Blinding*	Interventions	Outcomes	Statistical Analysis*	Results	Discussion	Funding*
Prendiville et al., 1988[52]	○	○	●	●	○	○	◐	○	○	○
Guyatt et al., 1993;[54] Guyatt et al., 1994[53]	○	○	◐	●	◐	●	●	●	○	○
Standards of Reporting Trials Group, 1994[55]	○	●	●	●	◐	●	●	●	●	○
Asilomar Working Group, 1996[56]	●	●	◐	◐	●	●	◐	●	●	●
Moher et al., 2001[57]	●	●	●	●	●	●	●	●	●	○
Clarke and Oxman, 1999[11]	○	●	●	●	◐	◐	◐	◐	○	○
Lohr and Carey, 1999[1]	○	●	◐	●	●	●	●	●	●	○

*Domains with at least one element with an empirically demonstrated basis (see Table 8).

According to their authors, five scales underwent rigorous scale development along with the calculation of inter-rater reliabilities;[34,35,37,38,40] the one component system was both rigorously developed and measured inter-rater reliability.[51] Several scales and checklists were given a Partial score for their development process;[14,27,30-32,48] three of these also reported inter-rater reliability.[30,32]

Evaluation of Systems According to Seven Domains Considered Informative for Study Quality

As noted above, we identified four empirically based quality domains. To these we added three domains derived from best practices—study population, interventions, and outcomes—that we regarded as important for evaluating the quality of RCTs.

Study Population. The most important element in the study population domain is the specification of inclusion and exclusion criteria for entry of participants in the trial. Although such criteria constrain the population being studied (thereby making the study less generalizable), they reduce heterogeneity among the persons being studied. In addition, the criteria reduce variability, which improves our certainty of claiming a treatment effect if one truly exists.

Interventions. Intervention is another important quality domain mainly for one of its elements—that the intervention be clearly defined. For reasons of reproducibility both within the study and for comparison with other studies, investigators ought to describe fully the intervention under study with respect to dose, timing, administration, or other factors. Paying careful attention to the details of an intervention also tends to reduce variability among the subjects, which also influences what can be said about the study outcome.

Outcomes. As important as it is to describe the intervention clearly, it is also critical to specify clearly the outcomes under study and how they are to be measured. Again, this is important for both reproducibility and to decrease variability.

Coverage of Seven Key Domains. We designated a set of high-performing scales or checklists pertaining to RCTs by assessing their coverage of the following seven domains: study population, randomization, blinding, interventions, outcomes, statistical analysis, and funding. As with the five systems identified for systematic reviews, we concluded that these eight systems for RCTs represent acceptable approaches that could be used today without major modifications (Table 17).

Two systems fully addressed all seven domains,[24,45] and six others addressed all but funding.[14,18,26,36,38,40] Two were rigorously developed.[38,40] We might assume that the rigorousness with which the instruments were developed is important for assessing quality, but this has not been tested. 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.

Table 17. Evaluation of Scales and Checklists for Randomized Controlled Trials, by Instrument and Seven Key Domains

	Domains						
Instrument	**Study Population**	**Random-ization***	**Blinding***	**Interventions**	**Outcomes**	**Statistical Analysis***	**Funding***
Chalmers et al., 1981[24]	●	●	●	●	●	●	●
Liberati et al., 1986[26]	●	●	●	●	●	●	○
Reisch et al., 1989[45]	●	●	●	●	●	●	●
van der Heijden and van der Windt,1996[36]	●	●	●	●	●	●	○
de Vet et al., 1997[18]	●	●	●	●	●	●	○
Sindhu et al., 1997[38]	●	●	●	●	●	●	○
Downs and Black,1998[40]	●	●	●	●	●	●	○
Harbour and Miller, 2001[14]	●	●	●	●	●	●	○

*Domains with at least one element with an empirically demonstrated basis (see Table 8).

Rating Systems for Observational Studies

Type of System or Instrument

Seventeen systems concerned observational studies (Grid 3). Of these, we categorized four as scales[31,32,40,69] and eight as checklists (Table 18)[12-14,45,47,49,50,70] We 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. In the presentation below, we discuss scales and checklists separately from guidance documents and EPC rating systems.

Evaluation of Systems According to Coverage of Empirical or Best Practices Domains

Empirical Domains. The nine domains used for assessing these systems (Grid 3) reflect general study design issues common to observational studies (see Table 9). Of these domains, two have empirical elements: comparability of subjects and funding or sponsorship. Because the funding domain had only one element, it was required to give that domain a full Yes. We did not require that systems address the empirical element, use of concurrent controls, to receive a full Yes grade for the comparability-of-subjects domain. With the exception of one checklist that received a Partial score,[70] all scales and checklists received a full Yes rating for the comparability-of-subjects domain. Only one checklist received a full Yes for the funding domain.[45]

Best Practices Domains. The remaining seven domains—study question, study population, exposure/intervention, outcomes, statistical analysis, results, and discussion—come from best practices criteria. These domains are typically evaluated when critiquing an observational study. Of the 12 scales and checklists in Table 18, half fully addressed study question;[14,31,32,40,45,70] the remainder did not address this domain at all.[12,13,47,49,50,69] Similarly, for the discussion domain, we gave Yes or Partial ratings to only seven instruments.[13,31,32,40,45,47,50] Many systems covered results as a study quality domain, either fully or in part.[13,14,31,32,40,45,49,50,70] We rated the study population, exposure/intervention, outcome measure, and statistical analysis domains as Yes or Partial on most of the scales and checklists we reviewed.

Of the 12 scales and checklists, three fully addressed all these best practices domains.[32,40,45] Five others addressed most of the seven domains to some degree: One omitted exposure/intervention,[31] two did not include study question,[13,50] and the remaining two missed the discussion domain.[14,70] The remaining four systems entirely omitted two or more of the seven domains.[12,47,49,60]

Table 18. Evaluation of Scales and Checklists for Observational Studies, by Specific Instrument and Nine Domains

Instrument	Domains								
	Study Question	Study Population	Comparability of Subjects*	Exposure/ Intervention	Outcome Measure	Statistical Analysis	Results	Discussion	Funding*
Reisch et al., 1989[45]	●	●	●	●	●	●	●	●	●
Spitzer et al., 1990[47]	○	●	●	●	●	●	○	●	○
Cho and Bero, 1994[31]	●	●	●	○	◐	●	●	●	○
Goodman et al., 1994[32]	●	●	●	●	●	●	●	●	○
Downs and Black, 1998[40]	●	●	●	●	●	●	●	●	○
Corrao et al., 1999[69]	○	●	●	●	◐	●	○	○	○
Ariens et al., 2000[70]	●	●	◐	●	●	●	●	○	○
Khan et al., 2000[12]	○	●	●	●	◐	●	○	○	○
New Zealand Guidelines, 2000[13]	○	●	●	◐	●	●	●	◐	○
NHMRC, 2000[49]	○	◐	●	◐	◐	◐	◐	○	○
Harbour and Miller, 2001[14]	●	●	●	●	●	●	●	○	○
Zaza et al., 2000[50]	○	●	●	●	●	●	●	●	○

*Domains with one element with an empirically demonstrated basis (see Table 9).

Guidance Documents and EPC Systems. Guidance documents pertinent to observational studies (Grid 3) were not developed as tools for assessing quality, but all of them included comparability of subjects and outcomes either partially or fully. Most also included study population, statistical analysis, and results. The two EPC rating systems for observational studies are the same as those used for RCTs but with minor modifications; they were evaluated using the observational quality domains. One EPC system fully covered seven of the nine domains;[60] it omitted study question and funding. The other EPC system covered four domains—fully addressing comparability of subjects and outcomes but only partially addressing statistical analysis and results.[64]

Evaluation of Systems According to Descriptive Attributes

Of the 12 scales or checklists relating to observational studies, six selected their quality items based on accepted criteria;[12,45,47,50,69,70] five systems used both accepted and empirical criteria for item selection;[13,14,32,40,49] and one scale was a modification of another system.[31] One system was rigorously developed and provided an estimate of inter-rater reliability.[40] Three others received a Partial score for rigorousness of development but reported inter-rater reliability as well.[31,32,70]

Evaluation of Systems According to Domains Considered Informative for Study Quality

To arrive at a set of high-performing scales or checklists pertaining to observational studies, we considered the following five domains: comparability of subjects, exposure/intervention, outcomes, statistical analysis, and funding or sponsorship. As before, we concluded that systems that cover these domains represent acceptable approaches for assessing the quality of observational studies. The inclusion of the two empirical domains is self-explanatory (comparability of subjects and funding or sponsorship); we explain below why we considered the following as critical domains.

Exposure or Intervention. Unlike RCTs where treatment is administered in a controlled fashion, exposure or treatment in observational studies is based on the clinical situation and may be subject to unknown biases. These biases may result from provider, patient, or health care system differences. Thus, a clear description of how the exposure definition was derived is critical for understanding the effects of that exposure on outcome.

Outcomes. Investigators need to supply a specific definition of outcome that is independent of exposure. The presence or absence of an outcome should be based on standardized criteria to reduce bias and enhance reproducibility.

Statistics and Analysis. Of the six elements in the statistical analysis domain, confounding assessment was considered essential for a full Yes rating. Observational studies are particularly subject to several biases; these include measurement bias (usually addressed by specific exposure and outcome definitions) and selection bias (typically addressed by ensuring the comparability among subjects and confounding assessment). We did not consider any of the remaining five statistical analysis elements—statistical tests, multiple comparisons, multivariate techniques, power calculations, and dose response assessments—as more important than any other when evaluating systems on this domain.

Coverage of Five Key Domains. Of the 12 scales and checklists we reviewed, all included comparability of subjects either fully or in part. Only one included funding or sponsorship and the other four domains we considered critical for observational studies.[45] Five additional systems fully included all four domains without funding or sponsorship (Table 19).[14,32,40,47,50] In choosing among these six systems for assessing study quality, one will have to evaluate which system is most appropriate for the task being undertaken, how long it takes to complete each system, and its ease of use. We were unable to evaluate these three instrument properties in the project.

Rating Systems for Diagnostic Studies

Type of System or Instrument

As discussed in Chapter 2, the domains that we used to compare systems for assessing the quality of diagnostic test studies are to be used in conjunction with those relevant for judging the quality of RCTs or observational studies. Thus, here we contrast systems on the basis of five domains—study population, adequate description of the test, appropriate reference standard, blinded comparison of test and reference, and avoidance of verification bias.

We identified 15 systems for assessing the quality of diagnostic studies. Seven are checklists (Grid 4);[12,14,49,75-78,111] of these, one is a test-specific instrument.[111] The remainder are guidance documents. In addition, three EPCs used systems to evaluate the quality of the diagnostic studies.[59,68,79,80] In the discussion below, we comment on the checklists separately from the guidance documents and EPC scales.

Evaluation of Systems According to Coverage of Empirical or Best Practices Domains

Empirical Domains. The five domains used for assessing these systems (Table 10 and Grid 4) reflect design issues specific to evaluating diagnostic tests. Three domains—study population, adequate description of the test, and avoidance of verification bias—have only a single, empirical element; the other two domains each contain two elements, one of which has an empirical base.

Of the generic checklists we reviewed (Table 20), three fully addressed all six domains.[49,77,78] Two systems dealt with four of the five domains either fully or in part.[12,14] One checklist, the oldest of those we reviewed, addressed only one domain fully—use of an appropriate reference standard—and partially addressed the blinded reference comparison domain.[75,76]

Table 19. Evaluation of Scales and Checklists for Observational Studies, by Instrument and Five Key Domains

Instrument	Domains				
	Comparability of Subjects	Exposure/ Intervention	Outcome Measure	Statistical Analysis	Funding
Reisch et al., 1989[45]	●	●	●	●	●
Spitzer et al, 1990[47]	●	●	●	●	○
Goodman et al., 1994[32]	●	●	●	●	○
Downs and Black, 1998[40]	●	●	●	●	○
Harbour and Miller, 2001[14]	●	●	●	●	○
Zaza et al., 2000[50]	●	●	●	●	○

Table 20. Evaluation of Scales and Checklists for Diagnostic Test Studies, by Specific Instrument and Five Domains

	Domains*				
Instrument	**Study Population**	**Adequate Description of Test**	**Appropriate Reference Standard**	**Blinded Comparison of Test and Reference**	**Avoidance of Verification Bias**
Sheps and Schechter, 1984;[75] Arroll et al., 1988[76]	○	○	●	◐	○
Cochrane Methods Working Group, 1996[77]	●	●	●	●	●
Lijmer et al., 1999[78]	●	●	●	●	●
Khan et al., 2000[12]	●	○	●	●	●
NHMRC, 2000[49]	●	●	●	●	●
Harbour and Miller, 2001[14]	◐	○	●	●	●

*All domains have at least one element based on empirical evidence (see Table 10).

Almost all of the nine guidance documents included all these domains. One omitted the avoidance of verification bias domain;[71] another omitted adequate description of the test.[6] Of the three EPC scales, two addressed all five domains either fully[80] or in part.[59,68] We gave the remaining EPC system a No for adequate description of the test under study, although the information about the test was likely to have been captured apart from the quality rating system.[79]

Evaluation of Systems According to Descriptive Attributes

The six checklists were all generic instruments. Two systems used accepted criteria for selecting their quality items;[75-77] three used both accepted and empirical criteria;[12,14,78] and one was a modification of another checklist.[49] We gave two checklists a Partial score for development rigor primarily because they involved some type of consensus process.[14,78] Only the oldest system we reviewed addressed inter-rater reliability. [75,76,111]

Evaluation of Systems According to Domains Considered Informative for Study Quality

We consider all five domains in Table 20 to be critical for judging the quality of diagnostic test reports. As noted there, 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]

Findings for Systems to Rate the Strength Of a Body of Evidence

Background

Chapter 2 describes the development of the Summary Strength of Evidence Grid (Grid 5A) and Overall Strength of Evidence Grid (Grid 5B) that appear in Appendix C. Table 11 outlines our domains—quality, quantity, and consistency—for grading the strength of a body of evidence and gave their definitions.

We reviewed 40 systems that addressed grading the strength of a body of evidence. In discussing these approaches, we focus on 34 systems identified from our searches and prior research separately from those developed by six EPCs. The non-EPC systems came from numerous international sources, with the earliest systems coming from Canada. Based on the affiliation of the lead author, they originated as follows: Canada (11), United States (10), United Kingdom (6), Australia/New Zealand (3), the Netherlands (3), and a multi-national consensus group (1).

Evaluation According to Domains and Elements

Grid 5A distills the detailed information in Grid 5B. We use the same rating scheme as we did for the quality grids: Yes (●, the instrument fully addressed the domain); No (○, it did not

address the domain at all); or Partial (◐, it addressed the domain to some extent). Our findings for each system are discussed below.

Quality. The quality domain included only one element that incorporated our definition of quality (cited in Chapter 1), which was based on methodologic rigor—that is, the extent to which bias was minimized. Although the 34 non-EPC systems we reviewed included study quality in some way—that is, quality was graded as fully or partially met—their definitions of quality varied. Many systems defined quality solely by study design, where meta-analyses of RCTs and RCTs in general received the highest quality grade;[87-89,91,112-121] we gave these systems a Partial score. Systems indicating that conduct of the study was incorporated into their definition of quality received a Yes score for this domain. [11-14,22,39,70,81-86,90,122-128]

Of the six EPC grading systems, five received a full Yes score for quality.[59,60,67,68,129] One EPC system received an NA (not available) for quality because published information about evidence levels for efficacy did not directly incorporate methodologic rigor.[66] However, we know that this EPC measures study quality as part of its evidence review process.

Quantity. We combined three elements—numbers of studies, sample size or power, and magnitude of effect—under the heading of "quantity." As indicated in Chapter 2, a full Yes for this domain required that two of the three elements be covered. The quantity domain included magnitude of effect with both numbers of studies and sample size because we felt that these three elements provide assurance that the identified finding is true.
Of the 34 non-EPC systems, 16 fully addressed quantity, [11,13,22,81-86,88,89,91,117,122,124,125,127] and 15 addressed quantity in part.[12,14,39,70,84,90,112-114,118,121,123,126,128] Three systems did not include magnitude of effect, number of studies, or sample size as part of their evidence grading scheme. [117,119,120]

All the EPC systems that assessed the strength of the evidence in their first evidence reports included at least one of the three attributes we required for quantity; five fully addressed this domain, [59,65-68] and one did so in part. [60]

Consistency. The consistency domain had only one element, but it could be met only if the body of evidence on a given topic itself comprised more than one study. This would typically occur in the development of systematic reviews, meta-analyses, and evidence reports for which numerous studies are reviewed to arrive at a summary finding. As indicated in Chapter 2, this domain is dichotomous; a Yes indicates that the system took consistency into account and a No indicates that the system appeared not to consider consistency in its view of the strength of evidence.
Of the 34 non-EPC systems, approximately half incorporated the consistency domain into their approach to grading strength of evidence.[11,12,14,39,70,81-91] Only one EPC system included this domain. [65]

Evaluation of Systems According to Three Domains That Address the Strength of the Evidence

Domains. As indicated in Table 21, 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. [65] 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 their literature review process but this domain is apparently not directly incorporated into the grading system.[66]

Domains, Publication Year, and Purpose of System. Whether the grading systems dealing with overall strength of evidence dealt with all three domains appeared to differ by year of publication. The more recent systems included, either fully or partially, all three domains more frequently than did the older systems. Of the 23 evidence grading systems that had been published before 2000, seven (30 percent) included quality, quantity, and consistency to some degree; the same was true for nine (82 percent) of the 11 systems published in 2000 or later. This wide disparity among the systems can be attributed to the consistency domain, which began to appear more frequently from 2000 onward.

As discussed above, many evidence grading systems came from the clinical practice guideline literature. Table 22 shows that, at least among the 34 non-EPC grading systems, whether the three domains were incorporated differed by year of publication and primary purpose (i.e., for guideline development *per se* or for evidence grading). The *nonguideline* systems seemingly tended to incorporate all three domains more than the guideline systems, and this trend appears to be increasing over time.

Table 21. Extent to Which 34 Non-EPC Strength of Evidence Grading Systems Incorporated Three Domains of Quality, Quantity, and Consistency

Number of Domains Addressed and Extent of Coverage	Number of Systems
All three domains	
Addressed fully	7[11,81-86]
Addressed fully or partially	9[12,14,39,70,87-91]
Two of three domains	
Addressed fully	5[13,22,122,124,125]
Addressed fully or partially	10[112-116,118,121,123,126-128]
One domain addressed fully or partially	3[117,119,120]

Table 22. Number of Non-EPC Systems to Grade Strength of Evidence, by Number of Domains Addressed, Primary Purpose for System Development, and Year of Publication

Number of Domains Addressed*	Guideline System		Non-Guideline System	
	Before 2000	After 2000	Before 2000	After 2000
3 domains addressed either partially or fully	3 [81,88,89]	5 [14,82,83,86,91]	4 [11,39,87,90]	4 [12,70,84,85]
<3 domains addressed either partially or fully	13 [112-116,118-123,125,126,128]	2 [13,22]	3 [117,121,126]	0

*For systems to grade strength of evidence, domains are quality, quantity, and consistency.

Evaluation of Systems According to Domains Considered Informative for Assessing the Strength of a Body of Evidence

Of the seven systems that fully addressed quality, quantity, and consistency,[11,81-86] four were used for developing guidelines or practice recommendations,[81-83,86] and the remaining three were used for promoting evidence-based health care.[11,84,85]

These seven systems are very different (Table 23). Three appear to provide hierarchical grading of bodies of evidence,[82,83,85] and a fourth provides this hierarchy as part of its recommendations language.[86] Whether a hierarchy is desired will depend on the purpose for which the evidence grading is being done. However, as a society, we are used to numerical grading systems for comparing educational attainment, restaurant cleanliness, or other qualities, and a hierarchical system to grade the strength of bodies of evidence would be well understood and received.

Table 23. Characteristics of Seven Systems to Grade Strength of Evidence

	Domain				
Source	**Quality**	**Quantity**	**Consistency**	**Strength of Evidence Grading System**	**Comments**
Gyorkos et al., 1994[81]	Validity of studies	Strength of association and precision of estimate	Variability in findings from independent studies	Overall assessment of level of evidence based on four elements: Validity of individual studies Strength of association between intervention and outcomes of interest Precision of the estimate of strength of association Variability in findings from independent studies of the same or similar interventions For each element a qualitative assessment of whether there is strong, moderate, or weak support for a causal association.	
Clarke and Oxman, 1999[11]	Based on hierarchy of research design, validity, and risk of bias	Magnitude of effect	Consistency of effect across studies	Questions to consider regarding the strength of inference about the effectiveness of an intervention in the context of a systematic review of clinical trials: How good is the quality of the included trials? How large and significant are the observed effects? How consistent are the effects across trials? Is there a clear dose-response relationship? Is there indirect evidence that supports the inference? Have other plausible competing explanations of the observed effects (e.g., bias or cointervention) been ruled out?	Other domains: 1. Dose-response relationship 2. Supporting indirect evidence 3. No other plausible explanation

Table 23. Characteristics of Seven Systems to Grade Strength of Evidence (continued)

	Domain				
Source	**Quality**	**Quantity**	**Consistency**	**Strength of Evidence Grading System**	**Comments**
Briss et al., 2000[82]	Threats to Validity: - Study description - Sampling - Measurement - Data analysis - Interpretation of results - Other Quality of Execution: - Good (0-1 threats) - Fair (2-4 threats) - Limited (5+ threats) Design suitability: Greatestconcurrent comparison groups and prospective measure-ment Moderateall retrospective designs or multiple pre or post measurements; no concurrent comparison group Leastsingle pre and post measurements; no concurrent comparison group or exposure and outcome measured in a single group at the same point in time.	Effect size -Sufficient -Large -Small Larger effect sizes (absolute or relative risk) are considered to represent stronger evidence of effective-ness than smaller effect sizes with judgments made on an individual basis	Consistency as yes or no.	Evidence of effectiveness is based on execution, design suitability, number of studies, consistency, and effect size Strong: Good and greatest, at least 2 studies consistent, sufficient Good/fair and great/moderate, at least 5 studies, consistent, sufficient Good/fair and any design, at least 5 studies, consistent, sufficient Sufficient Good and greatest, one study, consistency unknown, sufficient Good/fair and great/moderate, at least 3 studies, consistent, sufficient Good/fair and any design, at least 5 studies consistent, sufficient Expert opinion: sufficient effect size Insufficient: insufficient design, too few studies, inconsistent, small effect size	

Table 23. Characteristics of Seven Systems to Grade Strength of Evidence (continued)

	Domain				
Source	**Quality**	**Quantity**	**Consistency**	**Strength of Evidence Grading System**	**Comments**
Greer et al., 2000[83]	Strong design not defined but includes issues of bias and research flaws	System incorporates number of studies and adequacy of sample size	Incorporates consistency	Grade I: Evidence from studies of strong design; results are both clinically important and consistent with minor exceptions at most; results are free from serious doubts about generalizability, bias, and flaws in research design. Studies with negative results have sufficiently large samples to have adequate statistical power. II: Evidence from studies of strong design but there is some uncertainty due to inconsistencies or concern about generalizability, bias, research design flaws, or adequate sample size. Or, evidence consistent from studies of weaker designs. III: The evidence is from a limited number of studies of weaker design. Studies with strong design either haven't been done or are inconclusive. IV: Support solely from informed medical commentators based on clinical experience without substantiation from the published literature.	Does not require a systematic review of the literature—only six "important" research papers.

Table 23. Characteristics of Seven Systems to Grade Strength of Evidence (continued)

	Domain				
Source	**Quality**	**Quantity**	**Consistency**	**Strength of Evidence Grading System**	**Comments**
Guyatt et al., 2000[84]	Based on hierarchy of research design, with some attention to size and consistency of effect	Multiplicity of studies, with some attention to magnitude of treatment effects	Consistency of effect considered	Hierarchy of vidence for application to patient care: N of 1 randomized trial Systematic reviews of randomized trials Single randomized trials Systematic review of observational studies addressing patient-important outcomes Single observational studies addressing patient-important outcomes Physiologic studies Unsystematic clinical observations Authors also discuss a hierarchy of preprocessed evidence that can be used to guide the care of patients: Primary studies—by selecting studies that are both highly relevant and with study designs that minimize bias, permitting a high strength of inference Summaries—systematic reviews Synopses—of individual studies or systematic reviews Systems—practice guidelines, clinical pathways, or evidence-based textbook summaries	Evidence defined broadly as any empirical observation about the apparent relationship between events. "The hierarchy is not absolute. If treatment effects are sufficiently large and consistent, for instance, observational studies may provide more compelling evidence than most RCTs."

Table 23. Characteristics of Seven Systems to Grade Strength of Evidence (continued)

	Domain				
Source	**Quality**	**Quantity**	**Consistency**	**Strength of Evidence Grading System**	**Comments**
NHS Centre for Evidence Based Medicine, (http://cebm.jr2.ox.ac.uk) (Accessed12-2001) [85]	Based on hierarchy of research design with some attention to risk of bias	Multiplicity of studies, and precision of estimate	Homogeneity of studies considered	Criteria to rate levels of evidence vary by one of four areas under consideration (Therapy/ Prevention or Etiology/Harm; Prognosis Diagnosis and Economic nalysis). For example, for the first area (Therapy/ Prevention or Etiology/Harm) the levels of evidence are as follows: 1a: SR with homogeneity of RCTs 1b: Individual RCT with narrow 1c: All or none (this criteri met when all patients died the treatment becm available and now some survive or some died previously and now none die) 2a: with homogeneity of cohort studies 2b: Individual cohort study (including low quality RCT; e.g. <80% follow-up) 2c: “Outcomes” research 3a: SR with homogeneity of case-control studies 3b: Individual case-control study 4: Case-series and poor quality cohort and case-control studies 5: Expert opinion without explicit critical appraisal or based on physiology, bench research or “first principles.”	

Table 23. Characteristics of Seven Systems to Grade Strength of Evidence (continued)

	Domain				
Source	**Quality**	**Quantity**	**Consistency**	**Strength of Evidence Grading System**	**Comments**
Harris et al., 2001[86] (for the U.S. Preventive Services Task Force)	Based on hierarchy of research design and methodologic quality (good, fair, poor) within research design	Number of studies, see Consistency	Consistency Consistency is not required by the Task Force but if present, contributes to both coherence and quality of the body of evidence	Levels of evidence: I Evidence from at least one properly randomized controlled trial II-1 Well-designed controlled trial without randomization II-2 Well-designed cohort or case-control analytic studies, preferably from more than one center or group II-3 Multiple time series with or without the intervention (also includes dramatic results in uncontrolled experiments): III Opinions of respected authorities, based on clinical experience, descriptive studies, and case reports, or reports of expert committees • Aggregate internal validity is the degree to which the study(ies) provides valid evidence for the population and setting in which it was conducted. • Aggregate external validity is the extent to which the evidence is relevant and generalizable to the population and conditions of typical primary care practice. • Coherence/consistency	Other domains: Coherence Coherence implies that the evidence fits the underlying biologic model.

Chapter 4. Discussion

This chapter examines several discrete topics pertinent to the field of evidence-based practice and to efforts to develop rigorous reviews of the clinical and scientific knowledge on important health care issues. We first reflect on our data collection efforts for identifying the relevant literature because the challenges we encountered are instructive for others embarking on the development of systematic reviews and technology assessments. A second topic concerns how our results flow directly from how we conceptualized this project, giving due attention to the (perhaps conflicting) needs of policymakers, researchers, clinicians, and experts in evidence-based practice and to the implications of decisions about the empirical and epidemiologic analytic framework we used to structure our evaluations. Third, in earlier chapters we discussed our findings related to study quality independently of those for grading the strength of a body of evidence, and this strategy posed some issues that may influence our findings and conclusions. Finally, we offer our advice concerning directions for future research, noting that the challenges, gaps, and deficiencies in current rating or grading systems demand attention if the evidence-based practice field is to move forward with confidence and scientific rigor.

Data Collection Challenges

As noted in previous chapters, we identified 1,602 articles, reports, and other materials from our literature searches, web searches, referrals from our technical expert advisory group, and suggestions from independent peer reviewers of an earlier version of this report, and from a previous project conducted by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center (EPC) on behalf of the Agency for Healthcare Research and Quality (AHRQ). In the end, our formal literature searches were the least productive source of systems for this report. Of the more than 120 systems we 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 we could review, describe, and evaluate. Many articles from the search(es) 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.

We caution that those involved in evidence-based practice and research may not find it productive simply to search for quality rating schemes through standard (systematic) literature searches. This is one reason that we are comfortable with identifying (as in Chapter 3) a set of instruments or systems that meet reasonably rigorous standards for use in rating study quality. Little is to be gained by directing teams seeking to produce systematic reviews or technology assessments (or clinical practice guidelines) to initiate wholly new literatures searches in this area.

At the moment, we cannot provide concrete suggestions for how to search the literature on this topic most efficiently. Some advances must simply await expanded options for coding the peer-reviewed literature. Meanwhile, investigators wishing to build on our efforts might well consider tactics involving citation analysis and extensive contact with researchers and guideline developers to identify the systems they are presently using to assess the quality of studies in systematic reviews. In this regard, the efforts of at least some AHRQ-supported EPCs will be instructive.

Our literature search was most problematic for systems oriented toward grading the strength of a body of evidence. We found that the Medical Subject Headings (MeSH) terms were not very sensitive for identifying evidence grading systems. We attribute this phenomenon to the lag in development of MeSH terms specific for the evidence-based practice field.

To overcome this problem, we resorted to a text word search using "evidence," "strength," "rigor," "grading," and "ranking." This approach yielded nearly 700 articles, many of which reported the results of primary randomized controlled trials (RCTs). Our search yielded these articles because of a very common phrase: "no evidence that this treatment…" In other words, the trigger of the term "evidence" did not yield material concerned with grading the strength of a body of literature.

As a result, the systems we discussed in Chapter 3 (i.e., specifically those related to the entries in Grid 5 [Appendix C]) were identified primarily by reviewing the evidence grading schemes used by the authors of clinical guidelines and practice recommendations. Reliance on literature searches for finding instruments to assess bodies of evidence will likely prove disappointing, and we suggest that users, researchers, or policymakers wishing to explore this area today will need to rely on published materials cited in this report and contact with experts in the field for work in progress.

Conceptualization of the Project

Quality of Individual Articles

Types of Studies

We 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 individually. In the worst case, in fact, combining all such systems into a single evaluation framework risked nontrivial confusion and misleading conclusions, and we 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 scope of the project also dictated that we limit ourselves to the study designs most commonly encountered in clinical research. Other types of study designs do exist for which one might wish to evaluate study quality; among them are, for example, cost-effectiveness analysis and clinical prediction rules. However, the four designs we chose cover the vast majority of clinically relevant research and currently have a larger publication base from which to evaluate quality.

Domains and Elements Specific to Study Types

For these reasons, we developed separate assessments (as reflected in the grids in Appendix B and the tables in Chapter 3) to reflect this decision. Of necessity, each grid has its own set of domains for comparison. Grid 1 has 11 domains for evaluating the quality of systematic reviews, Grid 2 has 11 domains for RCTs, Grid 3 has nine domains for observational studies, and Grid 4 has five domains for studies evaluating diagnostic tests.

The domains for each type of study comprised one or more elements. Some were based directly on empirical results. As the literature highlighted in Appendix D shows, several empirical studies confirmed that bias *can* arise when certain design elements are not met. Thus, we considered these factors as critical elements for our study quality domains. Other domains or elements were based on best practices in the design and conduct of research studies. They 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 important implication of these points is that, because we chose the critical domains on which to judge systems, our results and recommendations are directly and inextricably linked to our definition of these domains (i.e., our conceptualization of the project). We believe that selecting such domains on the basis, mostly, of empirical work and, secondarily, on the grounds of long-standing best practices in epidemiology and clinical research is sound. Nonetheless, we note that other evaluators might opt to focus on different domains and, thus, come to different evaluations and conclusions.

For this reason, we emphasize that the "full" information on our assessments of all types of systems for the different study designs can be found in the grids in Appendix B, and we draw attention to both parts of those grids. (The first part provides our assessment of the degree to which the system dealt with all domains; the second part gives important descriptive information.) The tables in Chapter 3 then distill this information to highlight, for scales and checklists, the extent to which they cover all domains and then, the extent to which they cover domains we identified as crucial. We then focus on those systems that do an acceptable job of covering this latter set of domains.

In selecting among alternative systems, potential users of such systems may elect to return to the full grids to find information that they regard as critical to their decisionmaking. We also emphasize that the scope of our work did not permit our own application or testing of these instruments. Thus, at the moment, we must advise that potential users of any approaches identified in this report ought to give direct consideration to feasibility and ease of use and likely applicability to their own particular projects or topics.

Types of Systems

Although the project is focused on issues related to study quality, we also contrasted the systems in Grids 1-4 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 us to home in on scales and checklists as the more likely methods for rating articles that might be adopted more or less as is.

In some cases, guidance documents contained similar content but had not been devised for evaluative applications. We noted that a few of the guidance documents could, with relatively minimal effort, be reconstructed into a scale or checklist. In so doing, however, we would recommend that developers carry out some reliability and validity testing, as the lack of such testing for the scales and checklists we reviewed is a major gap in this field that ought not be perpetuated.

Strength of a Body of Evidence

Similarly, our grid concerning systems for grading the strength of bodies of evidence (Appendix C) is tied directly to our conceptual framework. As discussed in Chapter 2, we focused on three domains—quality, quantity, and consistency—because they combine important aspects of the collective design, conduct, and analysis of studies that address a given topic. Q*uality* here links back to the summation of the quality of individual articles. *Quantity* involves the magnitude of the estimated (observed) effect, the potential statistical power of the body of knowledge as reflected in the aggregate sizes of studies (i.e., their sample sizes), and the sheer number of studies bearing on the clinical or technology question under consideration. The accepted wisdom is that, all other things equal, a larger effect is better because a good deal of bias would have to be present to invalidate the likelihood of an association. Finally, *consistency* reflects the extent to which the results of included studies tell the same story and comport with known facts about the natural history of disease. These are well-established variables for characterizing how confidently we can conclude that a body of knowledge provides information on which clinicians or policymakers can act.

We did not include generalizability as a separate domain because we believed that our definition of consistency needed to focus only on concepts appropriate to grading the strength of a body of evidence. (In the evidence-based practice community, this idea is sometimes rendered as grading the strength of separate linkages in a comprehensive analytic framework or causal pathway.) In our view, generalizability (as it has typically been used in the clinical practice guideline arena) addresses whether the findings, aggregated across multiple studies, are relevant to particular populations, settings of care, types of clinicians, or other factors.

As we approached the tasks in this project, with the legislative mandate and AHRQ's history in mind, we concluded that our study ought to stop short of advising on the development or implementation of practice guidelines *per se.* Had we incorporated generalizability into our evaluative framework (as some peer reviewers suggested), our results and recommendations concerning systems for grading the strength of a body of evidence might have been very different.

Furthermore, including generalizability as a domain would have increased the complexity of our evaluations and added to the burden of applying them. Moreover, generalizability can be addressed only in the context of the clinical or technology question at hand—that is, to whom (e.g., patients, clinical specialties) or what settings is one interested in generalizing? In that sense, generalizability might be said to lie downstream of issues relating to study quality or strength of evidence, as we depicted in Figure 2. Finding generic grading systems that could deal clearly with different answers to that downstream question struck us as improbable, meaning that we might in the end have had fewer grading systems to suggest than we in fact identified in our results chapter.

Study Quality

Growth in Numbers of Systems

We identified at least three times as many scales and checklists for rating the quality of RCTs (n = 32) as we did for observational studies (n = 12), systematic reviews (n = 11), or diagnostic test studies (n = 6). We expect that ongoing methodological work addressing the quality of observational and diagnostic studies will over time affect both the number and the sophistication of these systems. Thus, our findings and conclusions with respect at least to observational and diagnostic studies may need to be readdressed once results from more methodological studies in these areas are available.

Development of Systems Appropriate for Observational Studies

As indicated in Appendix D, some empirical research is related to the design, conduct, and analysis of systematic reviews, RCTs, and studies evaluating diagnostic tests; much less information is presently available about the factors influencing the quality of observational studies. Many systems that we evaluated for observational studies (Grid 3) were ones that we also evaluated for RCTs (Grid 2). Reviewing the systems that apply to both types of study designs led us to conclude that the likely original intent of several of these systems was to evaluate the quality of RCTs and that the developers added questions to address observational studies as well.

Thus, abstracting information from and assessing these "one size fits all" systems against the two sets of relevant domains proved difficult (especially for the observational study grid). We see this as additional support for the view that a "single system" across all study types will not likely be achieved and, in fact, might be counterproductive.

The absence of systems specific to observational studies may be explained in part by the complexities involved in observational study design (a fact that can be appreciated from the flow diagram offered in Figure 1). RCTs improve the comparability between study and control groups using randomization to allocate treatments (preferably double-blinded randomization), and trialists attempt to maintain comparability of these groups by avoiding differential attrition or assessment.

By contrast, 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—we wonder whether nonmethodologically trained researchers can identify when potential selection bias or other biases more common with observational studies have occurred.

Longer or Shorter Instruments

When comparing across all the quality rating scales and checklists that we evaluated, we noted that the older ones tended to be most inclusive for the quality domains we chose to assess.[24,45] However, these systems also tended to be very long and potentially cumbersome to complete. As factors critical to good study design have been identified—that is, the empirical criteria we invoked in our assessments—we saw that the more recent systems are shorter and focus mainly on these empirical criteria for rating study quality.

Shorter instruments have the obvious advantage of brevity, and some data suggest that they will provide sufficient information on study quality. Jadad and colleagues reported that simply asking about three domains (randomization, blinding, and withdrawals [a form of attrition]) serves to differentiate between higher- and lower-quality RCTs that evaluate drug efficacy.[34] However, the Jadad scale is not applicable to study designs other than RCTs of therapies, and it is not very useful for health services interventions where randomization or double blinding cannot occur. The Jadad team also omitted elements such as allocation concealment and use of intention-to-treat statistical analysis. We judged that these two elements have an empirical basis, but we acknowledge that the information supporting them has emerged since the publication of their scale.

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, we are not convinced that shorter instruments *per se* will always be better, unless demonstrated in future empirical studies.

Reporting Guidelines

Several authors of the QUOROM and CONSORT statements served on our technical expert panel.[21,57] They strongly emphasized that such reporting guidelines are *not* to be used for assessing the quality of either RCTs or systematic reviews, respectively. We believe this is an appropriate caution, and so we considered these consensus works only as guidance documents in our review.

We applaud these consensus guidelines for reporting RCTs and systematic reviews. If these guidelines are used (and they are currently required by certain journals) they will lead to better reporting and two downstream benefits. First, this may diminish the unavoidable tension (when assessing study quality) between the actual study design, conduct, and analysis and the reporting of these study characteristics. Second, if researchers follow these guidelines when designing their studies, they are likely to have better designed studies that will then be more transparent when published.

Strength of a Body of Evidence

Interaction Among Domains

Our comparison of systems for assessing the strength of a body of evidence uses three domains (Grid 5). We did not try to unravel the interrelationships among quality, quantity, and consistency for this project. As the body of literature grows, additional studies (i.e., quantity) increase the likelihood of a large range of quality scores and heterogeneity with respect to population settings, outcomes measured, and results. When these factors are similar across studies, consistency (and thus, strength of evidence) is enhanced. When they are not, this heterogeneity will reduce consistency and presumably detract from the overall strength of the evidence. Alternatively, heterogeneity may provide clues that indicate important treatment differences across subpopulations under study.[130]

Conflict Among Domains When Bodies of Evidence Contain Different Types of Studies

Adding to the complexities of evaluating interactive domains for one type of study design is the challenge of evaluating a body of knowledge comprising observational and RCT data. As our peer reviewers pointed out, a contemporary case in point is the association between hormone replacement therapy (HRT) and cardiovascular risk.

Several observational studies, but only one large trial and two small RCTs, have examined the association between HRT and secondary prevention of cardiovascular disease for older women with preexisting heart disease.[131-133] In terms of quality, much of the observational work is considered good and the RCTs are considered very good. In terms of quantity, both the numbers of reports and individuals evaluated in these reports are high for observational studies and modest for 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. All three RCTs showed no evidence that hormone therapy was beneficial for women with established cardiovascular disease, and one RCT[133] found an increased risk of coronary events during the first year of HRT use.

Most experts would agree that RCTs minimize an important potential bias in the 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 we draw is that a system for grading 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.

Systems Related or Not Related to Development Of Clinical Practice Guidelines

Of the 34 non-EPC systems we evaluated for their performance in rating overall bodies of evidence, 23 addressed issues related to grading the strength of an evidence base for the development of clinical practice guidelines or treatment recommendations. The remaining 11 had not been derived directly from guideline development efforts *per se*. Interestingly, the first authors of all 11 of the non-guideline-derived systems are from outside the United States.[11,12,39,70,84,85,87,90,117,124,126]

Based on the results of this project, it appears that the only U.S. investigators who currently grade the strength of the evidence, *apart from those developing clinical practice guidelines or practice-related recommendations*, are those affiliated with AHRQ's EPCs. We believe a useful follow-on to the present study might be to evaluate more directly all the strength-of-evidence approaches now being used in guideline development as well as non-guideline development activities. Such an effort might well entail review of considerable collections of gray literature—for example, from the professional society's technical bulletins—rather than purely peer-reviewed publications.

Emerging Uses of Grading Systems

Two of the 11 non-guideline-derived systems graded the strength of the evidence for a systematic review of risk factors for back and neck pain.[70,90] Narrative and quantitative *systematic* reviews are typically done for therapies, preventive services, or diagnostic technologies—that is, to amass data that will inform clinical practice or reimbursement and coverage (policy) decisions. Traditional reviews are common for disease risk factors or health-related behaviors; evidence-based systematic reviews would be a likely next step as we move towards a greater reliance on evidence-based products for clinical or policy decisionmaking. Nonetheless, we are intrigued with this novel use of evidence grading for a systematic review on risk factors; it may foretell broader applications for systems of assessing study quality and evidence strength than has been seen to this point. Whether domains covered by extant rating and grading systems would need to be modified to take account of the types of research done to clarify risk factors is a matter of speculation and future research.

An example from the gray literature indicates that grading the strength of the evidence apart from the development of guidelines had been occurring even before the two risk evaluation studies[70,90] were published in the late 1990s. In 1994, the Institute of Medicine convened an expert panel to review the literature on the health effects of Agent Orange.[134] This team developed their own categorization system for grading the strength of this body of literature that also incorporated quality, quantity, and consistency.

As Guyatt and colleagues point out in their users' guides, summarizing the literature on treatment effects can (1) assist clinicians in treating their patients,[53,135] (2) help develop prevention strategies,[136] (3) resolve issues arising from conflicting studies of disease risk factors,[90] and (4) determine whether new treatments are worth their cost. Countries that have a national health service must identify ways to curb and prioritize health care spending, and many are turning to evidence-based practice to help them do so.

In the United States, we are beginning to see a rising emphasis on evidence-based practice and evidence-based policymaking. Like our foreign counterparts whose countries have national

health plans, we may begin to see policymakers in public programs such as Medicare and Medicaid placing a greater reliance on systematic reviews—and specifically systematic reviews that provide grades for the strength of evidence—documenting the benefits (and harms) of preventive, diagnostic, and therapeutic interventions relevant to those beneficiary populations. The same may prove to be true for administrative leaders of integrated health systems and managed care organizations. Certainly, study quality and evidence grading will be important issues when comparisons need to be made of diagnostic or therapeutic options for a given disorder using cost-effectiveness methodologies.

Limitations of the Research

Several limitations of the current research should be understood. The most important caveat is that the project team defined the quality and strength of the evidence domains for evaluation based on our review of the literature. We did so as objectively as possible and relied on well-respected work and the advice of our technical expert advisors. For our review of quality ratings, we included whatever quality domains the systems as a whole addressed, using as much detail as possible. However, our findings for all the grids are derived directly from our definitions and the way we structured this project.

Although our literature search was thorough and rigorous, it cannot be described as wholly systematic. Our two searches, one for identifying articles addressing study quality and the second for grading the strength of a body of evidence, dated from 1995 through June 2000. We searched only MEDLINE and restricted the articles to English language.

We did expand our search by viewing web sites known to contain publications prepared by groups from the United Kingdom, Canada, Australia, and New Zealand that focus on evidence-based medicine or guideline development. Moreover, our peer reviewers made suggestions for literature (e.g., on empirical bases for certain domains or for background and contextual materials) that had not surfaced as part of our formal literature searches. In addition, we did review several older articles that had been published as early as 1979. The more recent articles we identified as part of our literature search had cited the earlier publications as seminal pieces of work, and we would have been remiss in not including them in this project. All these additions, however, do make the formal data collection somewhat less “systematic” (but more comprehensive) than it might otherwise have been.

Finally, the time and resource constraints for this project led us to focus on generic study quality scales, checklists, and component systems. Although we included systems developed for narrow, specific clinical topics (e.g., pain; childhood leukemia; smoking-related diseases; drugs to treat alcohol dependence) that we encountered during the data collection phase, we did not actively seek them in our search. We see this gap as one that might profitably be filled by a second project to evaluate “specific” systems against the same types of criteria as applied here to “generic” instruments. Doing so would provide a more complete picture for potential users, investigators, or policymakers of the state of the science (and art) of rating the quality of articles and the strength of evidence today, and it will make clearer the contributions of those EPCs that have developed or adapted topic-specific approaches.

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

Rating Article Quality

In reviewing Grids 1-4 (Appendix B), we can see that many systems cover many of the domains that we considered generally informative for assessing study quality. However, we did not believe this range of information provided sufficient practical guidance for users who want to know, today, where to start. Thus, we condensed the information to identify systems that fully or at least partially addressed what we regarded as key domains, and these systems—largely scales and checklists—are the ones that appear in the tables of Chapter 3.

More specifically, we identified five systems for evaluating the quality of systematic reviews, eight for RCTs, six for observational studies, and three for studies of diagnostic tests (see Tables 14, 17, 19, and 21, respectively). Summing across these sets, we arrived at a total of 19 unduplicated systems that fully address our critical quality domains (with the exception of funding or sponsorship for several systems).[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.[14,40,45]

Based on this iterative analysis, we feel comfortable recommending that those who plan to incorporate study quality into a systematic review or evidence report 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 and the key methodological issues specific to the topic under study. We caution that systems ostensibly intended to be used to rate the quality of both RCTs and observational studies—what we 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.

We encourage those who will be incorporating study quality into a systematic review to examine many different quality instruments to determine which items will best suit their needs. We acknowledge that the resulting instrument will not be developed according to rigorous standards, but it will encompass domains that are important for the topic under evaluation. Other considerations for the selection and development of study quality systems include the topic to be reviewed, 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.

Rating Strength of Evidence

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. In addition, approaches for characterizing the strength of evidence seem to be getting longer or more complex with time. This trend stands in some contrast to that for systems related to assessing study quality, where the trend is for a reduction in the number of critical domains over time. This pattern may also reflect the fact that this effort is earlier on the development and diffusion curve.

Two other properties of these systems stand out. As discussed in Chapter 3, consistency has only recently become an integral part of the systems we reviewed in this area. We see this as a useful advance. Also continuing is the habit of using an older study design hierarchy to define

study quality as an element of grading overall strength of evidence. As recently noted in methodologic work done for the U.S. Preventive Services Task Force, however, reliance on such a hierarchy without consideration of the domains we have discussed throughout this report is increasingly seen as unacceptable. We would expect, therefore, that systems for grading strength of bodies of evidence will increasingly call for quality rating approaches like those identified above.

Table 23 in Chapter 3 provides the 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,[82-84] indicating that this is a rapidly evolving field.

As with the study quality 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 sufficient staff, time, and financial resources. Again, for users, researchers, and policymakers uncertain about which among these seven might best suit their needs, we suggest also applying descriptive information from Grid 5B in the decisionmaking.

EPC SYSTEMS

Although several EPCs used methods that met our criteria at least in part, these tended to be 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, we refer readers to the citations given earlier in this report.

Recommendations for Future Research

More than 30 empirical studies address design elements for systematic reviews, RCTs, observational studies, and studies to assess diagnostic tests (Appendix D). As can be inferred from our discussion throughout this report, insufficient information is available for identifying design elements proven to be critical for trials and other investigations (although this is less true for RCTs). Thus, as a general proposition, the information base for understanding how best to rate the quality of such studies remains incomplete. Until this research gap is bridged, those wishing to produce authoritative systematic reviews or technology assessments will be somewhat hindered in this aspect of their work.

In addition, most of the empirical work on study design issues at present pertains to systematic reviews and RCTs. Thus, more empirical research should be targeted to identify and resolve issues relevant to the quality of observational studies. Some information may arise shortly from the Cochrane Non-Randomised Studies Methods Group, which is drafting guidelines for using nonrandomized studies in Cochrane reviews. Our technical advisors also noted the work of the STARD (STAndards for Reporting Diagnostic accuracy) group, which will be providing a guideline for reporting of diagnostic test studies in the very near future.

The importance of inter-rater reliability for producing defensible systematic reviews and technology assessments should not be underestimated, especially in circumstances in which several reviewers (who may or may not be methodologically trained, as contrasted with clinically

trained) are contributing simultaneously to the review. Thus, another avenue for future research is to evaluate inter-rater reliability among the same and different quality systems as they may be applied for an evidence report or technology assessment of a given topic. This would be similar to the work done by Juni and colleagues, where they evaluated study quality using 25 different scales among publications addressing low molecular weight heparin and standard heparin post-surgery for prevention of deep vein thrombosis.[2]

Moreover, as implied above, rating study quality according to one of the "acceptable" systems that we have identified may be demonstrably easier and more reliable than grading strength of evidence according to systems examined for this project. For that reason, we emphasize the need for comparative work that uses several grading systems to evaluate the strength of the evidence on one topic as well as some reliability testing to determine whether several different reviews arrive at the same evidence grades.

We are encouraged that the U.S. Congress mandated this study from AHRQ in the first place. Nonetheless, our discussion in this chapter and earlier should make clear that a "one-shot" overview project could not and did not address all the significant issues in relating to methods or systems to assess health care research results.

We did not, for instance, give much attention to topic-specific approaches that may be somewhat more common in EPC work. In our judgment, one useful follow-up to the current project would assess whether the study quality grids that we developed are useful for discriminating among studies of varying quality—that is, as another set of study-specific quality systems. If they are useful for differentiation, a likely next step is to refine and test the systems further using typical instrument development techniques. Further valuable work would be to test the study quality grids against the instruments we have called out as meeting our final evaluation criteria. To assist such work, we have included (Appendix F) a reproduction of the data extraction forms used in this study.

Many of these systems have been developed abroad, and it seems clear that much of the activity in this area rests outside the United States. As evidence-based practice activities take even stronger hold in this country, through development of evidence reports, technology assessments, and clinical practice or health policy guidelines, we believe a more in-depth comparison and contrast might be made of how this work is done here and elsewhere. In particular, we believe that U.S. investigators should make strong efforts to ascertain what advances are taking place in the international community in efforts to develop systems for assessing study quality and evidence strength and to determine where these are relevant to the U.S. scene.

We noted the more common uses for such rating schemes as being for studies of therapies, preventive services, and diagnostic technologies. Further applications should be tested. As already mentioned, use of such approaches in studies of disease risk factors is one area of potentially fruitful research. Another is the extent to which existing approaches can be applied to the types of studies used to evaluate purely screening tests (as contrasted with tests used primarily for diagnosis). Finally, a significant emerging area concerns the efficacy or effectiveness of counseling interventions (whether for preventive or therapeutic purposes); such studies are often far more complex, heterogeneous, or multi-faceted than typical RCTs or observational studies, and we are not at all certain that existing rating and grading methods will apply. Therefore, examining the utility of the systems identified in this report for these "less traditional" bodies of evidence will be important in the future.

Many experts in this field point to the appreciable lack of proven elements and domains in these types of assessment instruments. Perusal of the tables in Chapter 2 that define domains and elements will indicate the extent to which we needed to rely on accepted practices in health services, clinical, and epidemiological research to populate the criteria by which we evaluated systems. Thus, a key item for the research agenda lies simply in extending the empirical work on these systems. Such work would show what factors used in rating study quality, for example, actually make a difference in final scores for individual articles or a difference in how quality is judged for bodies of evidence as a whole. In addition, we discussed earlier the contrasts between short and long forms of these rating and grading systems. All other things equal, shorter will be better because of the reduced burden on evaluators. Nonetheless, some form of "psychometric testing" of shorter forms in terms of reliability, reproducibility, and validity needs to be done, either of the shortened instrument itself or against its parent instrument.

A broader concern is the need to clarify techniques to make systematic reviews and technology assessments more efficient and cost-effective. Although that is not directly a matter solely for rating study quality and evidence strength, the potential link is that better methods for those tasks might enable investigators and evidence-based practice experts to arrive more easily at reviews in which the nature and merit of the knowledge base is clear to all.

Finally, we encourage greater experimentation and collaboration between U.S. and international professionals in commissioning and conducting systematic reviews and technology assessments. The AHRQ EPC program, with one EPC in Canada, is a good start, and collaboration does exist between two AHRQ EPCs (at Research Triangle Institute-University of North Carolina and at Oregon Health Sciences University) and their work or the U.S. Preventive Services Task Force and the equivalent Task Force in Canada. Moreover, AHRQ EPCs do examine reviews from the Cochrane Collaboration review groups in amassing literature on given issues.

Nonetheless, having multiple groups around the globe commissioning exhaustive reviews on essentially the same clinical or technology topics has obvious inefficiencies. Collaboration on the refinement of quality rating and evidence strength grading systems is one appealing step toward decreasing duplication, and broader coordination of work in the evidence-based arena may be desirable. The issue of generalizability or applicability of the evidence will certainly arise, but the literature base will basically be the same for all but highly country-specific health interventions and technologies.

Summary and Conclusion

To answer significant questions posed to AHRQ by the U.S. Congress, we reviewed more than 30 empirical studies to determine the critical domains for addressing study quality in each of four study designs: systematic reviews, RCTs, observational studies, and studies of diagnostic tests. Regardless of when this work was done, either recently or as long as 20 years ago, many investigators included most of the quality rating domains that we chose to assess.

We identified and reviewed, abstracted data from, and summarized more than 100 sources of information for the current study. Applying evaluative criteria based on key domains to the systems reported on in these articles, we identified 19 study quality and seven strength-of-evidence grading systems that those conducting narrative or quantitative systematic reviews and technology assessment can use as starting points. In making this information available to the Congress and disseminating information about these generic systems and the project as a whole

more widely, AHRQ can meet the congressional expectations 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 other groups developing clinical practice guidelines and other health-related policy advice. We 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. Thus, we 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. 1991;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. Harris RP, Helfand M, Woolf SH, et al. Current methods of the U.S. Preventive Services Task Force: A review of the process. Am J Prev Med. 2001;20:21-35.

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.

94. Mulrow CD. The medical review article: state of the science. Ann Intern Med. 1987;106:485-488.

95. Clark HD, Wells GA, Huet C, et al. Assessing the quality of randomized trials: reliability of the Jadad scale. Control Clin Trials. 1999;20:448-452.

96. Hemminki E. Quality of reports of clinical trials submitted by the drug industry to the Finnish and Swedish control authorities. Eur J Clin Pharmacol. 1981;19:157-165.

97. Khan KS, Daya S, Jadad A. The importance of quality of primary studies in producing unbiased systematic reviews. Arch Intern Med. 1996;156:661-666.

98. Field MJ, Lohr KN, eds. Guidelines for Clinical Practice: From Development to Use. Institute of Medicine. Washington, D.C.: National Academy Press; 1992.

99. Lohr KN, Aaronson NK, Burnam MA, Patrick DL, Perrin EB, Roberts JS. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther. 1996;18:979-991.

100. Last JM. A Dictionary of Epidemiology. New York: Oxford University Press; 1995.

101. Moher D, Jadad A, Tugwell P. Assessing the quality of randomized controlled trials. Int J Technol Assess Health Care. 1996;12:195-208.

102. Olkin I. Statistical and theoretical considerations in meta-analysis. J Clin Epidemiol. 1995;48:133-147.

103. Tritchler D. Modelling study quality in meta-analysis. Stat Med. 1999;18:2135-2145.

104. Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. Am J Epidemiol. 1994;140:290-296.

105. Chalmers TC, Celano P, Sacks HS, Smith HJ. Bias in treatment assignment in controlled clinical trials. N Engl J Med. 1983;309:1358-1361.

106. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ. 1998;317:1185-1190.

107. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials. 1995;16:62-73.

108. Fox JP, Hall CE, Elveback LR. Epidemiology; Man and Disease. New York: Macmillan; 1970.

109. Hill AB. The environment and disease: Association or causation? Proc R Soc Med. 1965;58:295.

110. Stelfox HT, Chua G, O'Rourke K, Detsky AS. Conflict of interest in the debate over calcium-channel antagonists. N Engl J Med. 1998;338:101-106.

111. Hoffman RM, Kent DL, Deyo RA. Diagnostic accuracy and clinical utility of thermography for lumbar radiculopathy. A meta-analysis. Spine. 1991;16:623-628.

112. Canadian Task Force on the Periodic Health Examination. The periodic health examination. Can Med Assoc J. 1979;121:1193-1254.

113. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1989;95:2S-4S.

114. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1992;102:305S-311S.

115. Ogilvie RI, Burgess ED, Cusson JR, Feldman RD, Leiter LA, Myers MG. Report of the Canadian Hypertension Society Consensus Conference: 3. Pharmacologic treatment of essential hypertension. Can Med Assoc J. 1993;149:575-584.

116. Evans WK, Newman T, Graham I, et al. Lung cancer practice guidelines: lessons learned and issues addressed by the Ontario Lung Cancer Disease Site Group. J Clin Oncol. 1997;15:3049-3059.

117. Granados A, Jonsson E, Banta HD, et al. EUR-ASSESS Project Subgroup Report on Dissemination and Impact. Int J Technol Assess Health Care. 1997;13:220-286.

118. Bartlett JG, Breiman RF, Mandell LA, File TMJ. Community-acquired pneumonia in adults: guidelines for management. The Infectious Diseases Society of America. Clin Infect Dis. 1998;26:811-838.

119. Bril V, Allenby K, Midroni G, O'Connor PW, Vajsar J. IGIV in neurology—evidence and recommendations. Can J Neurol Sci. 1999;26:139-152.

120. Working Party for Guidelines for the Management of Heavy Menstrual Bleeding. An evidence-based guideline for the management of heavy menstrual bleeding. N Z Med J. 1999;112:174-177.

121. Shekelle PG, Woolf SH, Eccles M, Grimshaw J. Clinical guidelines: developing guidelines. BMJ. 1999;318:593-596.

122. U.S. Preventive Services Task Force. Guide to Clinical Preventive Services, 2nd Ed. Alexandria, Va.: International Medical Publishing, Inc.; 1996.

123. Gross PA, Barrett TL, Dellinger EP, et al. Purpose of quality standards for infectious diseases. Infectious Diseases Society of America. Clin Infect Dis. 1994;18:421.

124. Gray JAM; Evidence-Based Healthcare. London: Churchill Livingstone;1997 .

125. Djulbegovic B, Hadley T. Evaluating the quality of clinical guidelines. Linking decisions to medical evidence. Oncology. 1998 Nov;12:310-314.

126. Edwards AG, Russell IT, Stott NC. Signal versus noise in the evidence base for medicine: an alternative to hierarchies of evidence? Fam Pract. 1998;15:319-322.

127. Chesson ALJ, Wise M, Davila D, et al. Practice parameters for the treatment of restless legs syndrome and periodic limb movement disorder. An American Academy of Sleep Medicine Report. Standards of Practice Committee of the American Academy of Sleep Medicine. Sleep. 1999;22:961-968.

128. Wilkinson CP. Evidence-based medicine regarding the prevention of retinal detachment. Transactions Am Ophthalmol Society. 1999;97:397-406.

129. Garbutt JC, West SL, Carey TS, Lohr KN, Crews FT. Pharmacological treatment of alcohol dependence: a review of the evidence. JAMA. 1999:281:1318-1325.

130. Berlin JA, Rennie D. Measuring the quality of trials: the quality of quality scales. JAMA. 1999;282:1083-1085.

131. Herrington DM, Reboussin DM, Brosnihan KB, et al. Effects of estrogen replacement on the progression of coronary-artery atherosclerosis. N Engl J Med. 2000;343:522-529.

132. Angerer P, Stork S, Kothny W, Schmitt P, von Schacky C. Effect of oral postmenopausal hormone replacement on progression of atherosclerosis: a randomized, controlled trial. Arterioscler Thromb Vasc Biol. 2001;21:262-268.

133. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. JAMA. 1998;280:605-613.

134. Committee to Review the Health Effects in Vietnam Veterans of Exposure to Herbicides; Divison of Health Promotion and Disease Prevention, Institute of Medicine. Veterans and Agent Orange. Washington, D.C.: National Academy Press; 1994.

135. Dans AL, Dans LF, Guyatt GH, Richard S. Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. JAMA. 1998;279:545-549.

136. Barratt A, Irwig L, Glasziou P, et al. Users' guides to the medical literature: XVII. How to use guidelines and recommendations about screening. Evidence-Based Medicine Working Group. JAMA. 1999;281:2029-2034.

137. Sacks HS, Berrier J, Reitman D, Anocaon-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. N Engl J Med.

1987;316:450-455.

138. Victor N. “The challenge of meta-analysis”:discussion. Indications and contra-indications for meta-analysis. J Clin Epidemiol. 1995;48:5-8

139. Longnecker MP, Berlin JA, Orza MJ, Chalmers TC. A meta-analysis of alcohol consumption in relation to risk of breast cancer. JAMA. 1988;260:652-656.

140. Begg CB. Biases in the assessment of diagnostic tests. Stat Med. 1987;6:411-423.

141. Working group on methods for prognosis and decision making. Memorandum for the Evaluation of Diagnostic Measures. Journal of Clinical Chemistry and Clinical Biochemistry. 1990;28:873-879.

142. Pinson AG, Becker DM, Philbrick JT, Parekh JS. Technetium-99m-RBC venography in the diagnosis of deep venous thrombosis of the lower extremity: a systematic review of the literature. J Nucl Med. 1991;32:2324-2328.

143. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1994;271:389-391.

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

145. Bruns DE. Reporting Diagnostic Accuracy. Clinical Chemistry. 1997;43(11):2211.

146. Becker DM, Philbrick JT, Abbitt PL, Real-time ultrasonography for the diagnosis of lower extremity deep venous thrombosis. The wave of the future? Arch Intern Med. 1989;149:173-1734.

147. Levine C, Armstrong K, Chopra S, Estok R, Zhang S, Ross S. Diagnosis and Management of Breast Disease: A Systematic Review of the Literature. Rockville, Md.: Agency for Healthcare Research and Quality; 2000.

148. 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.

149. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ. 1994;309:1286-1291.

150. Simes RJ. Publication bias: the case for an international registry of clinical trials. J Clin Oncol. 1986;4:1529-1541.

151. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet. 1991;337:867-872.

152. Jeng GT, Scott JR, Burmeister LF. A comparison of meta-analytic results using literature vs individual patient data. Paternal cell immunization for recurrent miscarriage. JAMA. 1995;274:830-836.

153. Moher D, Fortin P, Jadad AR, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. Lancet. 1996;347:363-366.

154. Moher D, Pham, Klassen TP, et al. What contributions do languages other than English make on the results of meta-analyses? J Clin Epidemiol. 2000;53:964-972.

155. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. Control Clin Trials. 1998;19:159-166.

156. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Knipschild PG. Balneotherapy and quality assessment: interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment. J Clin Epidemiol. 1998;51:335-341.

157. Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. Lancet. 1997;350:185-186.

158. Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical

study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Controlled Clinical Trials. 1990;11:339-352.

159. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ. 1994;309:1351-1355.

160. Chalmers TC, Matta RJ, Smith H Jr, Kunzler AM. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. N Engl J Med. 1977;297:1091-1096.

161. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. JAMA. 1994;272:125-128.

162. Grimes DA, Schulz KF. Methodology citations and the quality of randomized controlled trials in obstetrics and gynecology. American Journal of Obstetrics & Gynecology. 1996;174:1312-1315.

163. Chene G, Morlat P, Leport C, et al. Intention-to-treat vs. on-treatment analyses of clinical trial data: experience from a study of pyrimethamine in the primary prophylaxis of toxoplasmosis in HIV-infected patients. ANRS 005/ACTG 154 Trial Group. Control Clin Trials. 1998;19:233-248.

164. Lachin JM. Statistical considerations in the intent-to-treat principle. Control Clin Trials. 2000;21:167-189.

165. Djulbegovic B, Lacevic M, Cantor A, et al. The uncertainty principle and industry-sponsored research. Lancet. 2000;356:635-638.

166. Dong BJ, Hauck WW, Gambertoglio JG, et al. Bioequivalence of generic and brand-name levothyroxine products in the treatment of hypothyroidism. JAMA. 1997;277:1205-1213.

167. Rennie D. Thyroid storm. JAMA. 1997;277:1238-1243.

168. Cho MK, Bero LA. The quality of drug studies published in symposium proceedings. Ann Intern Med. 1996;124:485-489.

169. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342:1887-1892.

170. Barnes DE, Bero LA. Scientific quality of original research articles on environmental tobacco smoke. Tob Control. 1997;6:19-26.

171. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med. 1992;117:135-140.

Appendixes

Appendix A: Approaches to Grading Quality and Rating Strength of Evidence Used by Evidence-based Practice Centers

Introduction

An important element of this project was to summarize how the 12 evidence-based practice centers (EPCs) supported by the Agency for Healthcare Research and Quality (AHRQ) rated individual study quality and graded the strength of a body of evidence for their various evidence reports and technology assessments. The initial step in gathering information was for the AHRQ EPC Program Officer to ask the EPCs, on behalf of the team from the Research Triangle Institute-University of North Carolina (RTI-UNC) EPC, to identify the methods they used in these steps, assuming they did them at all. To assist in this process, RTI-UNC EPC staff reviewed the methods sections and appendices of all published evidence reports done by the EPCs for relevant information and then included this information with the form from the AHRQ Program Officer (Exhibit A-1) that asked the EPCs to specify how they handled quality ratings and evidence strength grading for their initial, subsequent, and current evidence reports and technology assessments. Several EPCs chose to summarize their procedures for us in a memorandum. We compiled the information (see Tables A-1 and A-2) and incorporated it into the appropriate grids (Appendices B and C).

Findings

Of the 12 EPCs, 10 did formally evaluate quality of articles in some fashion. Those that did applied numerous different techniques (Table A-1), and some based their quality assessments on study design only. Those that formally evaluated quality and developed a quality score employed several key study design components either as part of their inclusion/exclusion criteria or as components in their meta-analyses.

EPCs used quality ratings in several different ways:

1. As a factor in sensitivity or meta-analyses (Blue Cross and Blue Shield Association, Johns Hopkins University, New England Medical Center, Duke University, University of California at San Francisco-Stanford, University of Texas at San Antonio);
2. Descriptively in the evidence tables, results, and/or discussion section of their evidence reports (ECRI, McMaster University, Oregon Health Sciences University, RAND-Southern California, RTI-UNC); and
3. As inclusion/exclusion criteria for the literature searches of their evidence reports (MetaWorks, Inc.).

The data provided in Table A-2 are based on the completed surveys we received from each of the EPCs. Little changed over time with respect to whether and how the EPCs rated study

quality. Five EPCs graded the strength of bodies of evidence in their first EPC projects, and the same five currently grade evidence strength.

Exhibit A-1. Information Requisition Letter to EPCs

July 7, 2000

Dear EPC Directors and staff:

Thank you so very much for participating in our phone call with all the EPCs on May 22. As was discussed during the call, the RTI-UNC EPC has a very exciting but somewhat daunting task ahead of them and they need as much help from their fellow EPCs as they can possibly get!

The RTI-UNC EPC has organized an absolutely wonderful expert panel for this project. They include:

- Doug Altman
- Lisa Bero
- Alan Garber
- Steven Goodman
- Jeremy Grimshaw
- Alejandro Jadad
- Joseph Lau
- David Moher
- Cynthia Mulrow
- Andy Oxman
- Paul Shekelle

As Sue West indicated on the call, she has already reviewed the published AHRQ evidence reports (ERs) to identify the rating scales and methodologies for grading the evidence that were used by each EPC for their first ER (please see attached spreadsheet indicating which reports were reviewed). If information was available on rating scales or grading classifications from your ER(s), we are including a copy of the specific pages from your report with this letter. Please review this attached information to make sure that it accurately reflects the procedures you used at that time. Also, several of you very graciously provided Kathleen Lohr with information for her earlier project on the issues involved in grading articles and evidence so you certainly do not need to re-send this to the RTI-UNC EPC!

As the spreadsheet indicates, all of the EPCs have been funded to develop additional ERs. Your procedures may have changed somewhat as you worked on subsequent ERs. We would appreciate if you would share your procedures and full documentation that indicates how you are currently rating the quality of studies and grading the evidence so that the RTI-UNC EPC can document this in their report to AHRQ.

This information can be sent directly to Sue West at the following address:

Suzanne L. West, Ph.D., M.P.H.
Cecil G. Sheps Center for Health Services Research
725 Airport Road CB# 7590

University of North Carolina
Chapel Hill, NC 27599-7590

Alternatively, you can email to Sue_West@med.unc.edu or fax to 919-966-5764.

If you have suggestions for other scales or grading schemes or people to contact for this information, please provide this to Sue as well.

In their first ER, Pharmacotherapy for Alcohol Dependence, the RTI-UNC EPC did grade the evidence for their 5 key questions. With this letter, we have included copied pages from their evidence report that give their procedures as an example of what is meant by "grading the evidence" in the context of the current project, "Systems to Rate the Strength of the Scientific Evidence."

If you have any questions or need further guidance regarding your contributions to the RTI-UNC project, please don't hesitate to call Sue at 919-843-7662. Because of the timeline for this project, it would be great if you could send your information to Sue West by Friday, July 21. In replying to Sue, please include this letter and check the appropriate box indicating which information you are sending (or not sending!) to UNC. We (AHRQ and the RTI-UNC EPC) really appreciate your assisting the RTI-UNC EPC with this project.

- 1997 ER (first ER) for AHRQ

Rating study quality

❑ Contained forms and description for rating the quality of individual studies

→ ❑ Rating and description is being sent to the RTI-UNC EPC

↘ ❑ This info is **not** available to send

❑ Did not contain information on rating the quality of individual studies

Grading the evidence for key questions

❑ Contained information on grading the evidence

→ ❑ This info is being sent to the RTI-UNC EPC

↘ ❑ This info is **not** available to send

❑ Did not contain information on grading the evidence

- **Subsequent** ERs for AHRQ or other funding sources

Rating study quality

❑ Contained forms and description for rating the quality of individual studies

→ ❑ Rating and description is being sent to the RTI-UNC EPC

↘ ❑ This info is **not** available to send

❑ Did not contain information on rating the quality of individual studies

Grading the evidence for key questions

❑ Contained information on grading the evidence

→ ❑ This info is being sent to the RTI-UNC EPC

↘ ❑ This info is **not** available to send

❑ Did not contain information on grading the evidence

- Are you currently:

Rating study quality?

❑ Yes
❑ No

Grading the evidence for key questions?

❑ Yes
❑ No

Thank you, in advance, for your help!
Sincerely,

Jacqueline Besteman

cc: Kathleen N. Lohr, PhD
Valerie King, MD
Suzanne L. West, PhD, MPH

encl: EPC-specific pages from first evidence report
Pages from RTI-UNC Alcohol Pharmacotherapies evidence report
Spreadsheet with all projects
4-page project summary

Table A-1. Quality Ratings in Initial Evidence Reports of Evidence-based Practice Centers

Topic and *Nominators and Partners* by EPC	How Was Quality Measured in the Report?	How Was Quality Used in the Report?
Blue Cross and Blue Shield Association Technology Evaluation Center (TEC)		
Relative Effectiveness and Cost-Effectiveness of Methods of Androgen Suppression Treatment in the Treatment of Advanced Prostatic Cancer *Health Care Financing Administration*	Assessed the quality of methods and reporting to determine whether the studies could be grouped into categories by grade of methodologic quality. Factors assessed included: Random sequence generation Blinding of randomization process during recruitment Blinding of investigator and patient to treatment Study withdrawals Intent to treat Power Compliance with treatment Description of treatment protocols Formal quality rating was not given, component approach provided on evidence tables.	Quality used for sensitivity analyses. Meta-analysis combined hazard ratios for studies of "high" quality but "high" was not defined.
Duke University		
Evaluation of cervical cytology *American College of Obstetricians and Gynecologists*	Quality criteria for diagnostic tests • Test and reference measured independently • Test compared to valid reference standard • Choice of patients for reference standard independent of test results • Sample selection addressed • Location of publication • Funding source Consensus on the seven quality points, weight determined by consensus and averaging	Evaluated the effect of study quality on summary effectiveness scores using individual components of the score, then using the total score both as a continuous and dichotomous (cutpoint 7).
ECRI		
Diagnosis and treatment of dysphagia/swallowing problems in elderly *Health Care Financing Administration*	Quality measured by study design.	Study design reported in evidence tables and discussed in the results section of the evidence report.

Table A-1. Quality Ratings in Initial Evidence Reports of Evidence-based Practice Centers (cont.)

Topic and *Nominators and*Partners by EPC	How Was Quality Measured in the Report?	How Was Quality Used in the Report?
Johns Hopkins University		
Evaluation and treatment of new onset atrial fibrillation, in the elderly *American*Academy of Family Physicians	22 questions, major domains listed below: Thoroughness of population description Bias and confounding (description of randomization and blinding Standard protocol, other therapies received Outcomes and follow-up Statistical quality and interpretation	The EPC noted that it would have used study quality in a sensitivity analysis but there were too few studies.
McMaster University		
Treatment of attention deficit/ hyperactivity disorder American Academy of Pediatrics, *American Psychiatric Association*	Quality was based on the Jadad scale for randomized controlled trials Randomization Blinding Withdrawals Industry support	Authors used quality to describe results and conclusions.
MetaWorks, Inc.		
Diagnosis of sleep apnea *Blue Cross/Blue Shield of Massachusetts, Sleep Disorder Center of Metro Toronto*	Diagnostic studies rated by Irwig instrument before data abstraction Random order of assignment Use of a gold standard Blinded reading of test and gold standard	Quality score ranged from 0-44; papers with a quality score of <16 were not abstracted.
New England Medical Center		
Diagnosis and treatment of acute bacterial rhinosinusitis *American Academy of Otolaryngology, American Academy of Family Practice, American Academy of Pediatrics, American College of Physicians*	Quality was based on the Jadad scale for randomized controlled trials Randomization Blinding Withdrawals	Quality was used for sensitivity measure in meta-analysis Use of a composite quality score Use of factor(s) that relate to systematic bias

Table A-1. Quality Ratings in Initial Evidence Reports of Evidence-based Practice Centers (cont.)

Topic and *Nominators and Partners* by EPC	How Was Quality Measured in the Report?	How Was Quality Used in the Report?
Oregon Health Sciences University		
Rehabilitation of persons with traumatic brain injury *National Institute of Child Health and Human Development*, *Brain Injury Association*	Levels of quality Class I: properly designed RCTs Class II a: RCTs with design flaws or multicenter or population-based longitudinal (cohort) studies b: Controlled trials that were not randomized, case-control studies, case series with adequate description of population, intervention, outcomes Class III: descriptive studies, expert opinion, case reports, clinical experience	Quality levels were used descriptively in the results and conclusions section of the report.
RAND-Southern California Evidence-based Practice Center		
Prevention and management of urinary complications in paralyzed persons *Paralyzed Veterans of America, American Association of Spinal Cord Injury Psychologists, American Congress of Rehabilitation Medicine, American Paraplegia Society, Association of Rehabilitation Nurses, Consortium for Spinal Cord Medicine*	Quality was based on the Jadad scale for randomized controlled trials Randomization Blinding Withdrawals Cohort studies: Comparability at baseline or whether adjustments made during analysis Masked measurement of outcomes and risk factors	Quality grades were reported in evidence tables.
Research Triangle Institute––University of North Carolina, Chapel Hill		
Pharmacotherapy for alcohol dependence *American Society of Addiction Medicine*	Quality rating score adapted from scoring for spinal cord clinical guideline.	Authors reported quality scores in evidence tables and used them descriptively for results and conclusions.
University of California at San Francisco/Stanford University		
Management of stable angina *American College of Cardiology/American Heart Association Task Force on Practice Guidelines/American College of Physicians*	Four indicators: Randomization Blinding Description of randomization method Mention of exclusions	Authors used quality ratings in subgroup analyses.

Table A-1. Quality Ratings in Initial Evidence Reports of Evidence-based Practice Centers (cont.)

Topic and *Nominators and Partners* by EPC	How Was Quality Measured in the Report?	How Was Quality Used in the Report?
University of Texas at San Antonio EPC		
Depression treatment with new drugs *National Institute of Mental Health, American Psychiatric Association, American Pharmaceutical Association* **Vermont Department of Mental Health/Mental Retardation** *Blue Cross/Blue Shield of Massachusetts, American College of Physicians, Kaiser Permanente of Northern California*	Internal validity used instead of quality Randomization (method and concealment) Blinding Co-interventions Dropouts	Authors used the dropout rate in meta-analysis looking at response rates.

Table A-2. Summary of EPC Approach to Rating Quality and Grading the Strength of the Evidence from 1997 to July 2000

EPC	Subsequent Evidence Reports for AHRQ or Others		Current Practice	
	Rating Study Quality	Grading the Evidence for Key Questions	Rating Study Quality	Grading the Evidence for Key Questions
Blue Cross and Blue Shield	●	O	●	O
Duke University	●	O	●	O
ECRI	O	O	O	O
Johns Hopkins University	●	●	●	●
McMaster University	●	O	●	O
MetaWorks, Inc.	●	●	●	●
New England Medical Center	●	●	●	●
Oregon Health Sciences University	●	●	●	●
Southern California Evidence-based Practice Center-RAND	●	O	●	O
RTI-UNC	●	●	●	●
UCSF-Stanford	◐	O	O	O
UT - San Antonio	●	O	●	O

Legend:

● Yes

◐ Partial

O No

Appendix B: Quality of Evidence Grids

Quality Grid 1A. Evaluation of Quality Rating Systems for Systematic Reviews

Instrument	Domains										
	Study Question	Search Strategy	Inclusion/ Exclusion	Interventions	Outcomes	Data Extraction	Study Quality/ Validity	Data Synthesis & Analysis	Results	Discussion	Funding/ Support
Oxman and Guyett, 1991[4]; Oxman et al., 1991[5]	●	●	◐	○	◐	◐	●	●	●	○	○
Irwig et al., 1994[6]	●	●	●	●	●	●	●	●	●	○	○
Oxman et al., 1994[15]	●	●	◐	○	◐	◐	●	●	●	○	○
Cook et al., 1995[16]	●	●	●	◐	●	●	●	●	●	●	○
Sacks et al., 1996[7]	●	●	●	●	●	●	●	●	●	◐	●
Auperin et al., 1997[8]	◐	●	●	●	●	●	◐	●	●	◐	●
Beck, 1997[9]	●	●	●	○	○	●	○	●	●	●	○
Cranney et al., 1997[17]	○	●	●	○	◐	●	◐	◐	●	○	○
de Vet et al., 1997[18]	●	●	●	●	○	●	●	●	●	○	○
Smith, 1997[10]	◐	●	●	◐	○	○	●	◐	○	◐	○
Barnes and Bero, 1998[3]	●	◐	●	◐	○	○	●	●	◐	◐	●
Pogue and Yusuf, 1998[19]	●	◐	●	●	◐	◐	◐	●	◐	◐	○
Sutton et al., 1998[20]	●	●	●	○	●	●	●	●	●	●	○
Clarke and Oxman, 1999[11]	●	●	●	○	○	○	●	●	●	●	○
Moher et al., 1999[21]	●	●	●	●	●	●	●	●	●	●	○
Khan et al., 2000[12]	●	●	●	●	●	●	●	●	●	●	○

Quality Grid 1A. Evaluation of Quality Rating Systems for Systematic Reviews

Instrument	Domains										
	Study Question	Search Strategy	Inclusion/ Exclusion	Interventions	Outcomes	Data Extraction	Study Quality/ Validity	Data Synthesis & Analysis	Results	Discussion	Funding/ Support
New Zealand Guidelines Group, 2000[13]	●	●	●	○	●	○	○	○	●	◐	○
NHMRC, 2000[22]	○	●	●	●	○	○	●	●	○	○	○
Harbour and Miller 2001[14]	●	●	◐	●	●	○	●	●	●	○	○
Stroup et al., 2000[23]	●	●	●	◐	●	●	●	●	●	●	●

Note: For complete reference information, see reference list

Quality Grid 1B. Description of Quality Rating Systems for Systematic Reviews

Instrument	Description of Instrument to Assess Study Quality						
	Generic or specific instrument	Type of instrument?	Quality concept discussed	Method used to select items	Rigorous development process	Inter-rater reliability reported	Instructions provided for instrument use?
Oxman and Guyatt 1991[4]; Oxman et al., 1991[5]	Generic	Checklist	No	Accepted	No	ICC=0.71 (95%CI:0.59-0.81)	Yes
Irwig et al., 1994[6]	Generic	Checklist	No	Accepted	No	No	Partial
Oxman et al., 1994[15]	Generic	Guidance	Partial	Accepted	No	No	Partial
Cook et al., 1995[16]	Generic	Guidance	Partial	Both	Partial	No	Yes
Sacks et al., 1996[7]	Generic	Checklist	No	Modified Sacks, et al., 1987[137]	No	No	Yes
Auperin et al., 1997[8]	Generic	Checklist	No	Modified Sacks, et al., 1987[137]	No	ICC = 0.89-0.96	Partial
Beck, 1997[9]	Generic	Checklist	No	Modified multiple sources	No	% Agreement 87-89%	No
Cranney et al., 1997[17]	Generic	Guidance	No	Modified Victor, 1995[138] and Cook, 1995[16]	Partial	No	No
de Vet et al., 1997[18]	Specific	Guidance	Partial	Accepted	No	No	Partial
Smith, 1997[10]	Generic	Checklist	No	Modified Mulrow 1987[94] and Oxman, et al., 1994[15]	No	No	Partial
Barnes and Bero, 1998[3]	Generic	Scale	Partial	Modified Oxman, et al., 1994[15]	No	r = 0.87	No
Pogue and Yusuf, 1998[19]	Generic	Guidance	Partial	Accepted	No	No	Partial
Sutton et al., 1998[20]	Generic	Guidance	Yes	Modified multiple sources	No	No	Partial
Clarke and Oxman, 1999[11]	Generic	Checklist	No	Both	Partial	No	Partial
Moher et al., 1999[21]	Generic	Guidance	Yes	Both	Partial	No	Partial
Khan et al., 2000[12]	Generic	Checklist	Yes	Both	No	No	Partial
New Zealand Guidelines Group, 2000[13]	Generic	Checklist	No	Both	No	No	Partial

Quality Grid 1B. Description of Quality Rating Systems for Systematic Reviews

Instrument	Description of Instrument to Assess Study Quality						
	Generic or specific instrument	Type of instrument?	Quality concept discussed	Method used to select items	Rigorous development process	Inter-rater reliability reported	Instructions provided for instrument use?
NHMRC, 2000[22]	Generic	Guidance	Yes	Modified Clarke and Oxman (1999)[11]	No	No	Partial
Harbour and Miller 2001[14]	Generic	Checklist	Yes	Both	Partial	No	Yes
Stroup et al, 2000[23]	Generic	Guidance	Partial	Both	No	No	Partial

Note: For complete reference information, see reference list

ICC = intraclass correlation coefficient
k = kappa
R = correlation coefficient

Quality Grid 2A. Evaluation of Quality Rating Systems for Randomized Controlled Trials

Instrument	Domains									
	Study Question	Study Population	Randomization	Blinding	Interventions	Outcomes	Statistical Analysis	Results	Discussion	Funding/ Support
Chalmers et al., 1981[24]	○	●	●	●	●	●	●	●	○	●
DerSimonian et al., 1982[43]	○	●	●	●	○	○	◐	◐	○	○
Evans and Pollock, 1985[25]	●	●	◐	●	●	●	◐	●	●	○
Liberati et al., 1986[26]	○	●	●	●	●	●	●	●	○	○
Poynard et al., 1987[44]	○	●	●	●	○	●	◐	○	○	○
Prendiville et al., 1988[52]	○	○	●	●	○	○	◐	○	○	○
Colditz et al., 1989[27]	○	◐	●	●	○	●	◐	●	◐	○
Gotzsche, 1989[28]	○	○	◐	●	●	●	●	●	◐	○
Reisch et al., 1989[45]	●	●	●	●	●	●	●	●	●	●
Imperiale and McCullough, 1990[46]	○	◐	◐	○	●	◐	○	○	○	○
Spitzer et al., 1990[47]	○	●	◐	◐	●	●	◐	◐	●	○
Kleijnen et al., 1991[29]	◐	◐	◐	◐	●	●	○	●	○	○
Detsky et al., 1992[30]	○	●	●	◐	●	●	○	●	○	○
Guyatt et al., 1993[54]; Guyatt et al., 1994[53]	○	○	◐	●	◐	●	●	●	○	○
Cho and Bero,1994[31]	●	●	●	●	○	◐	◐	●	●	○
Goodman et al., 1994[32]	●	●	◐	◐	●	●	●	●	●	○

Quality Grid 2A. Evaluation of Quality Rating Systems for Randomized Controlled Trials

Instrument	Domains									
	Study Question	Study Population	Randomization	Blinding	Interventions	Outcomes	Statistical Analysis	Results	Discussion	Funding/ Support
Standards of Reporting Trials Group, 1994[55]	○	●	●	●	◐	●	●	●	●	○
Fahey et al., 1995[33]	○	●	◐	◐	●	○	◐	○	○	○
Schulz et al., 1995[51]	○	○	●	●	○	○	●	○	○	○
Asilomar Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature, 1996[56]	●	●	◐	◐	●	●	◐	●	●	●
Moher et al., 2001[57]	●	●	●	●	●	●	●	●	●	○
Jadad et al., 1996[34]	○	○	●	●	○	○	◐	○	○	○
Khan et al., 1996[35]	○	○	●	●	○	○	◐	○	○	○
van der Heijden et al., 1996[36]	○	●	●	●	●	●	●	●	○	○
Bender and Halpern, 1997[37]	○	○	●	◐	○	○	◐	○	○	○
de Vet et al., 1997[18]	○	●	●	●	●	●	●	●	○	○
Sindhu et al., 1997[38]	●	●	●	●	●	●	●	◐	●	○
van Tulder et al., 1997[39]	○	◐	●	●	●	◐	●	●	○	○
Downs and Black, 1998[40]	●	●	●	●	●	●	●	●	○	○
Moher et al., 1998[41]	○	○	●	●	○	●	●	○	○	○
Verhagen et al., 1998[48]	○	◐	●	●	○	◐	◐	●	○	○

Quality Grid 2A. Evaluation of Quality Rating Systems for Randomized Controlled Trials

Instrument	Domains									
	Study Question	Study Population	Randomization	Blinding	Interventions	Outcomes	Statistical Analysis	Results	Discussion	Funding/ Support
Clarke and Oxman, 1999[11]	○	●	●	●	◐	◐	◐	◐	○	○
Lohr and Carey, 1999[1]	○	●	◐	●	●	●	●	●	●	○
Khan et al., 2000[12]	○	●	●	●	●	○	●	●	○	○
New Zealand Guidelines Group, 2000[13]	○	●	◐	●	●	●	●	●	◐	○
NHMRC, 2000[49]	○	○	●	●	○	●	●	○	○	○
Harbour and Miller 2001[14]	●	●	●	●	●	●	●	●	○	○
Turlik and Kushner, 2000[42]	○	●	●	●	◐	◐	●	●	○	○
Zaza et al., 2000[50]	○	●	○	●	●	●	◐	●	●	○
EPC Quality Assessments										
Aronson et al.,1999[58]	○	○	●	●	●	○	●	◐	○	○
Chestnut et al.,1999[60]	○	●	◐	◐	●	●	◐	●	●	○
Jadad et al.,1999[61]	○	●	●	●	●	●	◐	●	○	●
Heidenreich et al., 1999[62]	○	◐	●	●	○	○	○	○	○	●
Mulrow et al., 1999[63]	○	●	●	●	●	●	◐	●	○	●
Vickrey et al.,1999[64]	○	○	●	●	○	○	◐	○	○	○
West et al., 1999[65]	●	●	◐	●	●	●	●	●	●	○
McNamara et al., 2001[66]	○	●	●	●	●	●	●	●	○	○

Quality Grid 2A. Evaluation of Quality Rating Systems for Randomized Controlled Trials

Instrument	Domains									
	Study Question	Study Population	Randomization	Blinding	Interventions	Outcomes	Statistical Analysis	Results	Discussion	Funding/ Support
Ross et al., 2000[67]	○	○	●	●	○	○	◐	○	○	○
Goudas et al., 2000[68] Lau et al., 2000[59]	○	○	●	●	○	○	◐	◐	○	○

Note: For complete reference information, see reference

Quality Grid 2B. Description of Quality Rating Systems for Randomized Controlled Trials

Instrument	Description of Instrument to Assess Study Quality						
	Generic or specific instrument	Type of instrument?	Quality concept discussed	Method used to select items	Rigorous development process	Inter-rater reliability reported	Instructions provided for instrument use
Chalmers et al., 1981[24]	Generic	Scale	Yes	Accepted	No	No	Yes
DerSimonian et al., 1982[43]	Generic	Checklist	Partial	Accepted	No	% Agreement 51-82%	Partial
Evans and Pollock, 1985[25]	Generic	Scale	No	Accepted	No	No	Yes
Liberati et al., 1986[26]	Generic	Scale	Yes	Modified Chalmers et al., 1981[24]	No	No	Partial
Poynard et al., 1987[44]	Generic	Checklist	No	Modified Chalmers 1981[24]	No	No	No
Prendiville et al., 1988[52]	Generic	Guidance	Yes	Accepted	No	No	Yes
Colditz et al., 1989[27]	Generic	Scale	Yes	Modified DerSimonian et al., 1982[43]	Partial	No	Partial
Gotzsche, 1989[28]	Specific	Scale	No	Accepted	No	No	Yes
Reisch et al., 1989[45]	Generic	Checklist	Partial	Accepted	No	Partial	Yes
Imperiale et al., 1990[46]	Generic	Checklist	Yes	Accepted	No	k = 0.79	No
Spitzer et al., 1990[47]	Generic	Checklist	Partial	Accepted	No	No	No
Kleijnen et al., 1991[29]	Generic	Scale	Partial	Accepted	No	Partial	Partial
Detsky et al., 1992[30]	Generic	Scale	Yes	Accepted	Partial	ICC = 0.92	Partial
Guyatt et al., 1993[54]; Guyatt et al., 1994[53]	Generic	Guidance	No	Accepted	No	No	Partial
Cho and Bero, 1994[31]	Generic	Scale	Yes	Modified Spitzer et al., 1990[47]	Partial	r = 0.60 ± 0.13	Partial
Goodman et al., 1994[32]	Generic	Scale	Yes	Both	Partial	ICC = 0.25	Yes
Standards of Reporting Trials Group, 1994[55]	Generic	Guidance	Yes	Both	Partial	No	Yes
Fahey et al., 1995[33]	Generic	Scale	Partial	Modified Clarke and Oxman (1999)[11]	No	No	No

Quality Grid 2B. Description of Quality Rating Systems for Randomized Controlled Trials

Instrument	Description of Instrument to Assess Study Quality						
	Generic or specific instrument	Type of instrument?	Quality concept discussed	Method used to select items	Rigorous development process	Inter-rater reliability reported	Instructions provided for instrument use
Schulz et al., 1995[51]	Generic	Component	Yes	Accepted	Yes	Partial	Yes
Asilomar Working Group, 1996[56]	Generic	Guidance	No	Both	Partial	No	No
Moher et al., 2001[57]	Generic	Guidance	Yes	Both	Partial	No	Yes
Jadad et al., 1996[34]	Generic	Scale	Yes	Empiric	Yes	ICC = 0.66 (95% CI: 0.53-0.79)	Yes
Khan et al., 1996[35]	Generic	Scale	Yes	Modified Jadad et al., 1996[34]	Yes	k = 0.70-0.94	Yes
van der Heijden et al., 1996[36]	Specific	Scale	Partial	Accepted	No	No	Yes
Bender et al., 1997[37]	Generic	Scale	Partial	Modified Jadad et al., 1996[34]	Yes	ICC = 0.85	Partial
de Vet et al., 1997[18]	Generic	Scale	Partial	Accepted	No	No	No
Sindhu et al., 1997[38]	Generic	Scale	No	Both	Yes	R = 0.90-0.99	Partial
van Tulder et al., 1997[39]	Generic	Scale	No	Accepted	No	No	No
Downs and Black, 1998[40]	Generic	Scale	Partial	Both	Yes	r = 0.75	Yes
Moher et al., 1998[41]	Generic	Scale	Yes	Modified Jadad et al., 1996,[34] and Schulz et al., 1995[51]	Yes	No	Partial
Verhagen et al., 1998[48]	Generic	Checklist	Yes	Both	Partial	No	No
Clarke and Oxman, 1999[11]	Generic	Guidance	Yes	Both	No	No	Partial
Lohr and Carey, 1999[1]	Generic	Guidance	Yes	Accepted	No	No	No
Khan et al., 2000[12]	Generic	Checklist	Yes	Both	No	No	Partial
New Zealand Guidelines Group, 2000[13]	Generic	Checklist	No	Both	No	No	Partial
NHMRC, 2000[49]	Generic	Checklist	Yes	Both	No	No	No
Harbour and Miller 2001[14]	Generic	Checklist	Yes	Both	Partial	No	Yes
Turlik and Kushner, 2000[42]	Specific	Scale	No	Both	No	No	No

Quality Grid 2B. Description of Quality Rating Systems for Randomized Controlled Trials

Instrument	Description of Instrument to Assess Study Quality						
	Generic or specific instrument	**Type of instrument?**	**Quality concept discussed**	**Method used to select items**	**Rigorous development process**	**Inter-rater reliability reported**	**Instructions provided for instrument use**
Zaza et al., 2000[50]	Generic	Checklist	No	Accepted	No	% Agreement 65.2-85.6 % (Median= 79.5%)	Yes
EPC Quality Assessments							
Aronson et al., 1999[58]	Specific	Checklist	No	Both	No	No	Partial
Chestnut et al., 1999[60]	Specific	Checklist	No	Both	No	No	Partial
Jadad et al., 1999[61]	Specific	Scale	Yes	Both	No	No	Partial
Heidenreich et al., 1999[62]	Specific	Checklist	No	Both	No	No	No
Mulrow et al., 1999[63]	Specific	Scale	No	Both	No	No	Yes
Vickrey et al., 1999[64]	Generic	Scale	Yes	Empiric	Yes	No	Yes
West et al., 1999[65]	Specific	Scale	Yes	Accepted	Partial	k=0.66-0.88	Yes
McNamara et al., 2001[66]	Specific	Scale	Partial	Modified Detsky et al., (1992)[30] and Clarke and Oxman (1999)[11]	No	Partial	Partial
Ross et al., 2000[67]	Generic	Scale	Yes	Modified Jadad et al., 1996[34]	No	Partial	Yes
Goudas et al., 2000[68] Lau et al., 2000[59]	Generic	Component	Yes	Accepted	No	No	Partial

Note: For complete reference information, see reference list

ICC = intraclass correlation coefficient
k = Kappa
R= correlation coefficient

Quality Grid 3A. Evaluation of Quality Rating Systems for Observational Studies

Instrument	Domains								
	Study Question	Study Population	Comparability Of Subjects	Exposure/ Intervention	Outcome Measure	Statistical Analysis	Results	Discussion	Funding
Reisch et al., 1989[45]	●	●	●	●	●	●	●	●	●
Spitzer et al., 1990[47]	○	●	●	●	●	●	○	●	○
Carruthers et al., 1993[71]	○	○	◐	○	●	○	○	○	○
Cho and Bero, 1994[31]	●	●	●	○	◐	●	●	●	○
Goodman et al., 1994[32]	●	●	●	●	●	●	●	●	○
Laupacis et al., 1994[72]	○	●	◐	○	●	◐	●	○	○
Levine et al., 1994[73]	○	◐	◐	●	●	◐	●	○	○
Downs and Black, 1998[40]	●	●	●	●	●	●	●	●	○
Angelillo and Villari, 1999[74]	○	●	●	◐	●	●	●	○	○
Corrao et al., 1999[69]	○	●	●	●	◐	●	○	○	○
Lohr and Carey, 1999[1]	○	●	◐	●	●	●	●	●	○
Ariens et al., 2000[70]	●	●	◐	●	●	●	●	○	○
Khan et al., 2000[12]	○	●	●	●	◐	●	○	○	○
New Zealand Guidelines, 2000[13]	○	●	●	◐	●	●	●	◐	○
NHMRC, 2000[49]	○	◐	●	◐	◐	◐	◐	○	○
Harbour and Miller 2001[14]	●	●	●	●	●	●	●	○	○
Zaza et al.,	○	●	●	●	●	●	●	●	○

Quality Grid 3A. Evaluation of Quality Rating Systems for Observational Studies

Instrument	Domains								
	Study Question	Study Population	Comparability Of Subjects	Exposure/ Intervention	Outcome Measure	Statistical Analysis	Results	Discussion	Funding
2000[50]									
EPC Quality Assessments									
Chestnut et al., 1999[60]	○	●	●	●	●	●	●	●	○
Vickrey et al., 1999[64]	○	○	●	○	●	◐	◐	○	○

Note: For complete reference information, see reference

Quality Grid 3B. Description of Quality Rating Systems for Observational Studies

Instrument	Description of Instrument to Assess Study Quality						
	Generic or specific instrument	Type of instrument?	Quality concept discussed	Method used to select items	Rigorous development process	Inter-rater reliability reported	Instructions provided for instrument use?
Reisch et al., 1989[45]	Generic	Checklist	Partial	Accepted	No	Partial	Yes
Spitzer et al., 1990[47]	Generic	Checklist	Partial	Accepted	No	No	No
Carruthers et al., 1993[71]	Generic	Guidance	No	Accepted	No	No	No
Cho and Bero, 1994[31]	Generic	Scale	Yes	Modified Spitzer et al., 1990[47]	Partial	r = 0.60 ± 0.13	Yes
Goodman et al., 1994[32]	Generic	Scale	Yes	Both	Partial	ICC = 0.25	Yes
Laupacis et al., 1994[72]	Generic	Guidance	No	Accepted	No	No	Partial
Levine et al., 1994[73]	Generic	Guidance	No	Accepted	No	No	Partial
Downs et al., 1998[40]	Generic	Scale	Partial	Both	Yes	r = 0.75	Yes
Angelillo and Villari, 1999[74]	Specific	Guidance	Partial	Modified Chalmers et al., 1981,[24] and Longnecker, 1988[139]	No	No	No
Corrao et al., 1999[69]	Specific	Scale	No	Accepted	No	No	No
Lohr and Carey, 1999[1]	Generic	Guidance	Yes	Accepted	No	No	No
Ariens et al., 2000[70]	Specific	Checklist	Yes	Accepted	Partial	% Agreement between 2 reviewers = 84%	Partial
Khan et al., 2000[12]	Generic	Checklist	Yes	Accepted	No	No	Partial
New Zealand Guidelines Group, 2000[13]	Generic	Checklist	No	Both	No	No	Partial
NHMRC, 2000[49]	Generic	Checklist	Yes	Both	No	No	Partial

Quality Grid 3B. Description of Quality Rating Systems for Observational Studies

Instrument	Description of Instrument to Assess Study Quality						
	Generic or specific instrument	**Type of instrument?**	**Quality concept discussed**	**Method used to select items**	**Rigorous development process**	**Inter-rater reliability reported**	**Instructions provided for instrument use?**
Harbour and Miller 2001[14]	Generic	Checklist	Yes	Both	Partial	No	Yes
Zaza et al., 2000[50]	Generic	Checklist	No	Accepted	No	% Agreement 65.2-85.6% (Median = 79.5%)	Yes
EPC Quality Assessments							
Chestnut et al., 1999[60]	Specific	Checklist	No	Both	No	No	Partial
Vickrey et al., 1999[64]	Generic	Scale	No	Accepted	No	No	Partial

Note: For complete reference information, see reference list

ICC = intraclass correlation coefficient
k = Kappa
R = correlation coefficient

Quality Grid 4A. Evaluation of Quality Rating Systems for Diagnostic Studies

Instrument	Domains				
	Study Population	Adequate Description of Test	Appropriate Reference Standard	Blinded Comparison of Test and Reference	Avoidance of Verification Bias
Sheps and Schechter, 1984[75]; Arroll et al., 1988[76]	○	○	●	◐	○
Begg, 1987[140]	◐	●	●	●	●
Working Group on methods for prognosis and decision making, 1990[141]	●	●	●	●	●
Hoffman et al., 1991[111]	●	●	●	●	●
Pinson et al., 1991[142]	●	●	●	●	●
Carruthers et al., 1993[71]	●	●	●	●	○
Jaeschke et al., 1994[143]	◐	●	●	●	●
Irwig et al., 1994[6]	◐	○	●	●	●
Reid et al., 1995[144]	●	●	●	●	●
Cochrane Methods Working Group, 1996[77]	●	●	●	●	●
Bruns ,1997[145]	●	●	●	●	●
Lijmer et al., 1999[78]	●	●	●	●	●
Khan et al., 2000[12]	●	○	●	●	●
NHMRC, 2000[49]	●	●	●	●	●
Harbour and Miller, 2001[14]	◐	○	●	●	●
EPC Quality Assessments					
McCrory et al., 1999[79]	●	○	●	●	●
Ross et al., 1999[80]	●	●	●	●	●

Quality Grid 4A. Evaluation of Quality Rating Systems for Diagnostic Studies

Instrument	Domains				
	Study Population	Adequate Description of Test	Appropriate Reference Standard	Blinded Comparison of Test and Reference	Avoidance of Verification Bias
Goudas et al., 2000;[68] Lau et al., 2000[59]	◐	●	●	●	◐

Note: For complete reference information, see reference

Quality Grid 4B. Description of Quality Rating Systems for Diagnostic Studies

Instrument	Description of Instrument to Assess Study Quality						
	Generic or specific instrument	Type of instrument?	Quality concept discussed	Method used to select items	Rigorous development process	Inter-rater reliability reported	Instructions provided for instrument use?
Sheps and Schechter, 1984[75]; Arroll et al., 1988[76]	Generic	Checklist	No	Accepted	No	k = 0.81-1.0	Partial
Begg, 1987[140]	Generic	Guidance	Partial	Accepted	No	No	Partial
Working Group, 1990[141]	Generic	Guidance	No	Accepted	Partial	No	Partial
Hoffman et al., 1991[111]	Specific	Checklist	Partial	Based on multiple other systems	No	k = 0.61	Yes
Pinson et al., 1991[142]	Specific	Guidance	No	Modified Becker, 1989[146]	No	No	Yes
Carruthers et al., 1993[71]	Generic	Guidance	No	Accepted	No	No	no
Jaeschke et al., 1994[143]	Generic	Guidance	No	Accepted	No	No	Partial
Irwig et al., 1994[6]	Generic	Guidance	Yes	Accepted	No	No	Partial
Reid et al., 1995[144]	Generic	Guidance	Yes	Accepted	No	No	Yes
Cochrane Methods Working Group, 1996[77]	Generic	Checklist	Partial	Accepted	No	No	Yes
Bruns, 1997[145]	Generic	Guidance	Partial	Accepted	Partial	No	Partial
Lijmer et al., 1999[78]	Generic	Checklist	Yes	Both	Partial	No	Partial
Khan et al., 2000[12]	Generic	Checklist	Yes	Both	No	No	Partial
NHMRC, 2000[49]	Generic	Checklist	Yes	Modified Clarke and Oxman,1999[11]	No	No	Partial
Harbour and Miller 2001[14]	Generic	Checklist	Yes	Both	Partial	No	Yes
EPC Quality Assessments							
McCrory et al., 1999[79]	Generic	Scale	No	Accepted	No	No	Partial
Ross et al., 1999[80]	Specific	Scale	Yes	Accepted	Partial	No	Partial
Goudas et al., 2000[68], Lau et al., 2000[59]	Generic	Component	Yes	Accepted	No	No	Partial

Note: For complete reference information, see reference

ICC = intraclass correlation coefficient
k = Kappa
R= correlation coefficient

Appendix C: Strength of Eviendce Grids Acronyms

ACRONYM	DESCRIPTION
CC	Case-control study
CI	Confidence interval
EB	Evidence-based
MA	Meta-analysis
N	Number
NA	Not available
NNT	Number needed to treat
OR	Odds ratio
RCT	Randomized controlled trial
SR	Systematic review

Grid 5A. Summary Evaluation of Systems for Grading by Three Domains			
	Domain		
	Quality	**Quantity**	**Consistency**
Source			
Canadian Task Force, 1979[112]	◐	◐	○
Anonymous, 1981[87]	◐	◐	●
Cook et al.,1992[114]; Sackett, 1989[113]	◐	◐	○
U.S. Preventive Services Task Force, 1996[122]	●	●	○
Ogilvie et al., 1993[115]	◐	●	○
Gross et al., 1994[123]	●	◐	○
Gyorkos et al., 1994[81]	●	●	●
Guyatt et al., 1998[88]	◐	●	●
Guyatt et al., 1995[89]	◐	●	●
Evans et al.,1997[116]	◐	◐	○
Granados et al., 1997[117]	◐	○	○
Gray, 1997[124]	●	●	○
van Tulder et al., 1997[39]	●	◐	●
Bartlett et al., 1998[118]	◐	◐	○
Djulbegovic and Hadley, 1998[125]	●	●	○
Edwards et al., 1998[126]	◐	◐	○
Bril et al., 1999[119]	◐	○	○
Chesson et al., 1999[127]	●	●	○
Clarke and Oxman, 1999[11]	●	●	●
Hoogendoorn et al., 1999[90]	●	◐	●
Working Party, 1999[120]	◐	○	○
Shekelle et al.,1999[121]	◐	◐	○
Wilkinson, 1999[128]	●	◐	○

Grid 5A. Summary Evaluation of Systems for Grading by Three Domains (Cont'd)			
	Domain		
	Quality	Quantity	Consistency
Source			
Ariens et al., 2000[70]	●	◐	●
Briss et al., 2000[82]	●	●	●
Greer et al., 2000[83]	●	●	●
Guyatt et al., 2000[84]	●	●	●
Khan et al., 2000[12]	●	◐	●
NHMRC, 2000[49]	●	●	○
NHS, 2001[85]	●	●	●
New Zealand Guidelines Group, 2000[13]	●	●	○
Sackett et al., 2000[91]	◐	●	●
Harbour and Miller, 2001[14]	●	◐	●
Harris et al., 2001[86]	●	●	●
EPC Quality Assessments			
Chestnut et al., 1999[60]	●	◐	○
West et al., 1999[65]	●	●	●
McNamara et al., 1999[66]	○	●	○
Ross et al., 2000[67]	●	●	○
Levine et al., 2000[147]	●	●	○
Goudas et al., 2000,[68] and Lau et al., 2000[59]	●	●	○

Note: For complete reference information, see reference list

Legend:

● Yes

◐ Partial

○ No information

Grid 5B. Overall Description of Systems to Grade Strength of Evidence

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Canadian Task Force on the Periodic Health Examination, 1979,1997[11,12] (This is the methodology section from the Web site www.ctfphc.org/Methodologyaccessed on 1-24-01)	Based on hierarchy of research design	Number of studies	NA		Quality of published evidence hierarchy: I Evidence from at least 1 properly randomized controlled trial II-1 Evidence from well-designed controlled trials without randomization II-2 Evidence from well-designed cohort or case-control analytic studies, preferably from more than 1 center or research group II-3 Evidence from comparisons between times or places with or without the intervention. Dramatic results in uncontrolled experiments could also be included here. III Opinions of respected authorities, based on clinical experience, descriptive studies or reports of expert committees.	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Anonymous, CMAJ, 1981[87]	Best evidence from RCTs (See Question 1 in comments column.)	Effect size and gradient (Question 2 in comments column.)	Consistency of association (Question 3 in comments column.)	See Questions 4–9 under comments section, which address issues of temporality, dose-response, epidemiologic and biologic sensibility and analogy.	Rates the relative importance of the various factors influencing a decision about causality listed in the comments section on a nine point scale from ++++ (supporting causation) to -- - - - (causation rejected), with 0 marking the point where causation is not affected Test / Test consistent with causation / Test neutral or inconclusive / Test opposes causation Human experiment: ++++ / – – – / – – – – Strength from: RCT: ++++ / – – – / – – – – Cohort: +++ / – – / – – – Case-Co: + / 0 / – Consistency: +++ / – – / – – – Temporality: ++ / – – / – – – – Gradient: ++ / – / – – Epidem. sense: ++ / – / – – Biologic sense: + / 0 / – Specificity: + / 0 / – Analogy: + / 0 / 0	Uses a series of 9 questions (diagnostic tests) for interpreting evidence of causation: 1. Is there evidence from true experiments in humans (i.e., is there evidence from RCTs)? 2. Is the association strong (i.e., how large is the measure of effect)? 3. Is the association consistent from study to study? 4. Is the temporal relationship correct? 5. Is there a dose-response relationship? 6. Does the association make epidemiologic sense? 7. Does the association make biologic sense? 8. Is the association specific? 9. Is the association analogous to a previously proven causal association?

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Cook et al., 1992[114] Sackett et al., 1989[113]	Based on hierarchy of research design	Sample size	NA		Levels of evidence: I Randomized trials with low false-positive (∀) and low false-negative (∃) errors II Randomized trials with high false-positive (∀) and/or high false-negative (∃) errors III Nonrandomized concurrent cohort comparisons between contemporaneous patients who did and did not receive therapy IV Nonrandomized historical cohort comparisons between current patients who did receive therapy and former patients who did not V Case series without controls	
U. S. Preventive Services Task Force,1996[122]	Based on hierarchy of research design, conduct of study, and risk of bias	Number of studies and statistical power to measure differences in effect	NA		Levels of evidence: I Evidence from at least one properly randomized controlled trial II-1 Well-designed controlled trial without randomization II-2 Well-designed cohort or CC analytic studies, preferably from more than one center or group II-3 Multiple time series with or without the intervention (also includes dramatic results in uncontrolled experiments) III Opinions of respected authorities, descriptive studies and case reports, reports of expert committees	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Ogilvie et al., 1993[115]	Based on hierarchy of research design	Considers statistical significance, sample size, and power	NA		Levels of evidence for rating studies of treatment: I An RCT that demonstrates a statistically significant difference in at least one important outcome. Alternatively, if the difference is not statistically significant, an RCT of adequate sample size to exclude a 25% difference in relative risk with 80% power, given the observed results. II An RCT that does not meet the level I criteria III A non-randomized trial with contemporaneous controls selected by some systematic method (i.e., not selected by perceived suitability for one of the treatment options for individual patients). Alternatively, subgroup analysis of an RCT. IV A before-after study or case series (of at least 10 patients) with historical controls or controls drawn from other studies. V Case series (at least 10 patients) without controls VI Case report (fewer than 10 patients)	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Gross et al., 1994[123]	Based on hierarchy of research design and conduct of study	Number of studies	NA		Levels of evidence: I Evidence from at least 1 properly randomized controlled trial II Evidence from at least 1 well-designed clinical trial without randomization, from cohort or case-controlled experiments (preferably from more than one center), multiple time-series studies, or dramatic results from uncontrolled studies III Opinions of the panel or respected authorities based on clinical judgment or descriptive studies IV Other: — Unanimous agreement — General, not unanimous	
Gyorkos et al., 1994[81]	Validity of studies	Strength of association and precision of estimate	Variability in findings from independent studies		Overall assessment of level of evidence based on four elements: 1 Validity of individual studies 2 Strength of association between intervention and outcomes of interest 3 Precision of the estimate of strength of association 4 Variability in findings from independent studies of the same or similar interventions For each element a qualitative assessment of whether there is strong, moderate or weak support for a causal association.	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Guyatt et al., 1998[88]	Based on hierarchy of research design	Multiplicity of studies, and precision of estimate relative to a treatment threshold	Consistency of study result considered		Levels of Evidence: Level I *(Grade A)* I Results come from a single RCT in which the lower limit of the CI for the treatment effect exceeds the minimal clinically important benefit I+ Results come from a meta-analysis of RCTs in which the treatment effects from individual studies are consistent, and the lower limit of the CI for the treatment effect exceeds the minimal clinically important benefit I- Results come from a meta-analysis of RCTs in which the treatment effects from individual studies are widely disparate, but the lower limit of the CI for the treatment effect still exceeds the minimal clinically important benefit	From Fifth ACCP Consensus Conference on Antithrombotic Therapy "...the more balanced the trade-off between benefits and risks the greater the influence of individual patient values in decision-making."

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
					Level II *(Grade B)* II Results come from a single RCT in which the CI for the treatment effect overlaps the minimal clinically important benefit II+ Results come from a meta-analysis of RCTs in which the treatment effects from individual studies are consistent and the CI for the treatment effect overlaps the minimal clinically benefit II- Results comefrom a meta-analysis of RCTs in which the treatment effects from individual studies are widely disparate and the CI for the treatment effect overlaps the minimal clinically important benefit Level III *(Grade C)* III Results come from nonrandomized concurrent cohort studies Level IV *(Grade C)* IV Results come from nonrandomized historic cohort studies Level V *(Grade C)* V Results come from case series	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Guyatt et al., 1995[89]	Based on hierarchy of research design	Number of studies and precision of estimate	Heterogeneity of studies and differences in estimates of effect considered		A1 RCTs, no heterogeneity, CIs all on one side of the threshold NNT A2 RCTS, no heterogeneity, CIs overlap threshold NNT B1 RCTs, heterogeneity, CIs all on one side of the threshold NNT B2 RCTs, heterogeneity, CIs overlap threshold NNT C1 Observational studies, CIs all on one side of the threshold NNT C2 Observational studies, CIs overlap threshold NNT	Authors define 2 criteria for what constitutes important heterogeneity among RCTs: 1. Difference in the estimate of RR reduction between the two most disparate studies is greater than 20%, and 2. The difference between the boundaries of the CIs between the two most disparate studies is greater than 5%. Their system uses 3 components to grade recommendations: strength of evidence, whether the impact of treatment warrants use and how effective the treatment is relative to a threshold number needed to treat (NNT). The grades range from A1 to C2 and are based on these three factors. For this strength of evidence grid we have abstracted only the A through C grades, which pertain most strongly to strength of evidence.

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Evans et al., 1997[116]	Based on hierarchy of research design	Adequacy of sample size to minimize false-positive or false-negative conclusions	NA		Levels: I Randomized controlled trials that are big enough to be either: - Positive with small risk of false-positive conclusions - Negative with small risk of false-negative conclusions - Meta-analysis II Randomized controlled trials that are too small, so that they show either: - Positive trends that are not statistically significant, with big risks of false-positive conclusions - No impressive trends but large risks of false-negative conclusions III Formal comparisons with non-randomized contemporaneous controls IV Formal comparisons with historic controls V Case-series	
Granados et al., 1997[117]	Based on hierarchy of research design	NA	NA		Level/strength of evidence upon which to base conclusions about the dissemination of technology assessments: 1 Strong; based on empirical evidence, including experimental and quasi-experimental data 2 Moderate; clear consensus among committee members 3 Weak; insufficient evidence, but viewed as worth considering by committee members	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Gray, 1997[124]	Based on hierarchy of research design and execution	Number of studies and power	NA		Strength of evidence: 1 Strong evidence from at least one systematic review of multiple, well-designed randomized controlled trials 2 Strong evidence from at least one properly designed randomized controlled trial of appropriate size 3 Evidence from well-designed trials without randomization, single group pre-post, cohort, time series, or matched case-control studies 4 Evidence from well-designed non-experimental studies from more than one center or research group 5 Opinions of respected authorities, based on clinical evidence, descriptive studies or reports of expert committees	
van Tulder et al., 1997[39]	Based on hierarchy of research design and conduct of study	Number of studies	Contradictory findings rated as Level 4 evidence		Levels of evidence: 1 Strong evidence—multiple relevant, high quality RCTs 2 Moderate evidence—one relevant, high quality RCT and one or more relevant, low quality RCTs 3 Limited evidence—one relevant, high quality RCT or multiple relevant, low quality RCTs 4 No evidence—only one relevant, low quality study, no relevant RCTs or contradictory outcomes	Based on rating system used for the U.S. Clinical Practice Guideline for Acute Low Back Problems in Adults.

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Bartlett et al., 1998[118]	Based on hierarchy of research design	Number of studies	NA		Evidence grade: I Evidence from at least one RCT II Evidence from at least one well-designed clinical trial without randomization III Evidence from opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees	
Djulbegovic et al., 1998[125]	Based on hierarchy of research design	Based partially on error rate <u>Error rate</u>: Low: acceptable false-positive rate 5%; acceptable false-negative rate 20% Intermediate: false-positive rate cannot be computed Highest: hints of efficacy only	NA		Levels: I Well-designed prospective randomized controlled trials with a low error rate.* II A single arm, prospective study, intermediate error rate.* III Retrospective/anecdotal date with the highest error rate.* *See quality column for definition of error rate.	Considers error rate and research design for grading the strength of the evidence

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Edwards et al., 1998[126]	Methodological quality	Effect size	NA		Concept of Signal-to-Noise Ratio: The authors suggest that the weight of evidence be assessed by comparing "signal" to "noise." Signal depends largely on effect size, but is assessed in the light of relevance and applicability to a particular situation. Noise refers to design deficiencies or methodological weaknesses.	
Bril et al., 1999[119]	Based on hierarchy of research design	NA	NA		A+ Randomized controlled, double-blind trials A Randomized controlled trials B Controlled trials C Open trials D Retrospective audits E Case-reports, expert opinion	
Chesson et al., 1999[127]	Based on hierarchy of research design	Considers alpha and beta error	NA		I Randomized well-designed trials with low alpha and low beta errors II Randomized trials with high beta errors III Nonrandomized controlled or concurrent cohort studies IV Nonrandomized historical cohort studies V Case series	Adapted from Sackett[113]

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Clarke and Oxman (Cochrane Collaboration Handbook) 1999[11]	Based on hierarchy of research design, validity and risk of bias	Magnitude of effect	Consistency of effect across studies	1 Dose-response relationship 2 Supporting indirect evidence 3 No other plausible explanation	Questions to consider regarding the strength of inference about the effectiveness of an intervention in the context of a systematic review of clinical trials: • How good is the quality of the included trials? • How large and significant are the observed effects? • How consistent are the effects across trials? • Is there a clear dose-response relationship? • Is there indirect evidence that supports the inference? • Have other plausible competing explanations of the observed effects (e.g., bias or cointervention) been ruled out?	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Hoogendoorn et al., 1999[90]	High quality: methodological quality score ≥50% of the maximum score Low quality: methodological quality score <50% of the maximum score	Number of studies	Inconsistent: if <75% of the available studies reported the same conclusion		Evidence based on quality, number, and the outcome of studies: Strong — provided by generally consistent findings in multiple high-quality studies Moderate — generally consistent findings in 1 high-quality study and 1 low-quality study, or in multiple low-quality studies No evidence — only 1 study available or inconsistent findings in multiple studies.	
Working Party for Guidelines for the Management of Heavy Menstrual Bleeding, 1999[120]	Based on hierarchy of research design	NA	NA		Grade A — Evidence based on randomized controlled trials Grade B — Evidence based on robust experimental or observational studies Grade C — Evidence based on more limited evidence but the advice relies on expert opinion and has the endorsement of respected authorities	Adapted from the National Health Service, United Kingdom. Grading is for quality of evidence and is based primarily on research design.

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Shekelle et al., 1999[121]	Based on hierarchy of research design	Multiplicity of studies	NA		Category of evidence: IA Evidence from meta-analysis of RCTs IB Evidence from at least one randomized controlled trial IIA Evidence from at least one controlled study without randomization IIB Evidence from at least one other type of quasi-experimental study III Evidence from non-experimental descriptive studies, such as comparative studies, correlation studies, and case-control studies IV Evidence from expert committee reports or opinions or clinical experience of respected authorities, or both	
Wilkinson, 1999[128]	Based on design, execution, and analysis	Typically one study	NA		Levels: I Strong evidence, i.e., study design addressed the issue in question, study was performed in the population of interest, and was executed to ensure accurate and reliable data with appropriate statistical analysis II Substantial evidence, i.e., study had some of the Level I attributes but not all of the attributes III Consensus of expert opinion without Level I or Level II evidence.	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Ariens et al., 2000[70]	Based on hierarchy of research design	Multiplicity of studies	Consistency of findings		Levels of evidence: 1 Strong evidence: consistent findings in multiple high-quality cohort or case-referent studies 2 Moderate evidence: consistent findings in multiple cohort or case-referent studies, of which only one study was high quality 3 Some evidence: findings of one cohort or case-referent study, or consistent findings in multiple cross sectional studies, of which at least one study was high quality 4 Inconclusive evidence: all other cases (i.e., consistent findings in multiple low quality cross-sectional studies, or inconsistent findings in multiple studies)	Applied to the question of physical risk factors for neck pain, hence only observational studies available for analysis.

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Briss et al., 2000[82]	Threats to validity: - study description - sampling - measurement - data analysis -interpretation of results - other Quality of execution: Good (0-1 threats) Fair (2-4 threats) Limited (5+ threats) Design suitability: *Greatest*- concurrent comparison groups and prospective measurement *Moderate*- all retrospective designs or multiple pre or post measurements; no concurrent comparison group *Least*- single pre and post-measurements; no concurrent comparison group or exposure and outcome measured in a single group at the same point in time.	Effect size - sufficient - large - small Larger effect sizes (absolute or relative risk) are considered to represent stronger evidence of effectiveness than smaller effect sizes with judgments made on an individual basis	Consistency as yes or no.		Evidence of effectiveness is based on execution, design suitability, number of studies, consistency, and effect size Strong: Good and greatest,* at least 2 studies, consistent, sufficient Good/fair and great/mod,* at least 5 studies consistent, sufficient Good/fair* and any design, at least 5 studies consistent, sufficient Sufficient: Good and greatest,* one study, consistency unknown, sufficient Good/fair and great/mod,* at least 3 studies consistent, sufficient Good/fair* and any design, at least 5 studies consistent, sufficient Expert opinion: sufficient effect size Insufficient: insufficient design, too few studies, inconsistent, small effect size *See description under Quality column	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Greer et al., 2000[83]	Strong design not defined but includes issues of bias and research flaws	System incorporates number of studies and adequacy of sample size	Incorporates consistency		Grade: I Evidence from studies of strong design; results are both clinically important and consistent with minor exceptions at most; results are free from serious doubts about generalizabiltiy, bias, and flaws in research design. Studies with negative results have sufficiently larded samples to have adequate statistical power. II Evidence from studies of strong design but there is some uncertainty due to inconsistencies or concern about generalizabiltiy, bias, research design flaws, or adequate sample size. Or, evidence consistent from studies of weaker designs. III The evidence is from a limited number of studies of weaker design. Studies with strong design either haven't been done or are inconclusive. IV Support solely from informed medical commentators based on clinical experience without substantiation from the published literature.	Does not require a systematic review of the literature—only six "important" research papers.

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Guyatt et al., 2000[84]	Based on hierarchy of research design, with some attention to size and consistency of effect	Multiplicity of studies, with some attention to magnitude of treatment effects	Consistency of effect considered		Hierarchy of evidence for application to patient care: • N of 1 randomized trial • Systematic reviews of randomized trials • Single randomized trials • Systematic review of observational studies addressing patient-important outcomes • Single observational studies addressing patient-important outcomes • Physiologic studies • Unsystematic clinical observations Authors also discuss a hierarchy of preprocessed evidence that can be used to guide the care of patients: • Primary studies—by selecting studies that are both highly relevant and with study designs that minimize bias, permitting a high strength of inference • Summaries—systematic reviews • Synopses—of individual studies or systematic reviews • Systems—practice guidelines, clinical pathways, or EB textbook summaries	Evidence defined broadly as any empirical observation about the apparent relationship between events. "The hierarchy is not absolute. If treatment effects are sufficiently large and consistent, for instance, observational studies may provide more compelling evidence than most RCTs."
Khan et al., 2000[12]	Based on hierarchy of research design	Sample size and power for providing precise estimates	Referred to as heterogeneity among studies		Level of evidence: 1 High quality experimental studies without heterogeneity and with precise results 2/3 Low quality experimental studies, high quality controlled observational studies 4 Low quality controlled observational studies, case series 5 Expert opinion	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
National Health and Medical Research Council, 2000[22]	Did the study design eliminate bias? How well were the studies done? Were appropriate and relevant outcomes measured?	How big was the effect? Does the p-value or confidence interval reasonably exclude chance?	NA		Levels of evidence: I Evidence obtained from a SR of all relevant RCTs II Evidence obtained from at least one properly designed RCT III-1 Evidence obtained from well-designed pseudorandomized controlled trial III-2 Evidence obtained from comparative studies (including SR of such studies) with concurrent controls and allocation not randomized, cohort studies, case-control studies, in interrupted time series with a control group III-3 Evidence obtained from comparative studies with historical control, two or more single arm studies, or interrupted time series without a parallel control group IV Evidence obtained from case series, either post-test or pretest/post-test	In the guidelines process NHMRC asks other questions to assess the evidence: Were appropriate and relevant outcomes measured? Was the effect clinically important? Levels of evidence now exclude expert opinion and consensus from an expert committee, although such forms of evidence were admitted in the 1995 guidance.

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
NHS Centre for Evidence Based Medicine, (http://cebm.jr2.ox.ac.uk) (accessed 12-2001)[85]	Based on hierarchy of research design with some attention to risk of bias	Multiplicity of studies, and precision of estimate	Homogeneity of studies considered		Criteria to rate levels of evidence vary by one of four areas under consideration (Therapy/Prevention or Etiology/Harm; Prognosis; Diagnosis; and Economic analysis). For example, for the first area (Therapy/Prevention or Etiology/Harm) the levels of evidence are as follows: 1A SR with homogeneity of RCTs 1B Individual RCT with narrow CI 1C All or none (this criteria met when all patients died prior to the treatment becoming available and now some survive or some died previously and now none die) 2A SR with homogeneity of cohort studies 2B Individual cohort study (including low quality RCT; e.g. <80% follow-up) 2C "Outcomes" research 3A SR with homogeneity of case-control studies 3B Individual case-control study 4 Case-series and poor quality cohort and case-control studies 5 Expert opinion without explicit critical appraisal or based on physiology, bench research or "first principles."	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
New Zealand Guidelines Group, 2000[13]	Based on hierarchy of research design and validity	Multiplicity of studies, magnitude of effect and range of certainty	NA		Evidence: 1 Randomized controlled trials 2 Non-randomized controlled studies 3 Non-experimental designs: — Cohort studies — Case control 4 Case series 5 Expert opinion	Evidence grades 1 through 5 appear to be based on study type, but text also discusses the importance of evaluating the actual study validity. This system is designed for application to questions of effectiveness. They distinguish between grading evidence and critical appraisal—for purposes of this summary we've merged these functions.
Sackett et al., 2000[91]	Based on hierarchy of research design	Considers narrowness of CI which relates to sample size and extent of follow-up	Homogeneity exhibited in systematic reviews		Level of evidence: 1A SR (with homogeneity) of RCTs 1B Individual RCT (with narrow CI) 1C All or none—prior to availability of new therapy, all died, now with therapy some survive 2A SR (with homogeneity of cohort studies 2B Individual cohort study (including low-quality RCT; e.g. <80% follow-up 2C "Outcomes" research 3A SR (with homogeneity of case-control studies 3B Individual case-control study 4 Case series (and poor-quality cohort and case-control studies) 5 Expert opinion without explicit critical appraisal or based on physiology, bench research or "first principles"	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Harbour and Miller, 2001[14]	Based on hierarchy of research design and risk of bias in conduct of study	Multiplicity of studies	Consistency of evidence considered in guidelines development process		SIGN's 1 through 4 level of evidence grading system is based on type of study, quality of study and risk of bias: 1++ High quality meta-anal, SR of RCTs or RCTs with very low risk of bias 1+ Well conducted meta-anal, SR of RCTs or RCTs with low risk of bias 1- Meta-analysis, SR of RCTs or RCTs with high risk of bias 2++ High quality SR of CC or cohort studies with very low risk of confounding or bias, and a high probability that relationship is causal 2+ Well conducted CC or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal 2- CC or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal 3 Non-analytic studies (e.g. case series) 4 Expert opinion	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Harris et al., 2001[86] Work for the U.S. Preventive Services Task Force	Based on hierarchy of research design and methodologic quality (good, fair, poor) within research design	Magnitude of effect (Numbers of studies or sizes of study samples are typically discussed by the USPSTF as part of this domain)	Consistency (Consistency is not required by the Task Force but if present, contributes to both coherence and quality of the body of evidence)	Coherence (Coherence implies that the evidence fits the underlying biologic model.)	Levels of evidence: I Evidence from at least one properly randomized controlled trial II-1 Well-designed controlled trial without randomization II-2 Well-designed cohort or CC analytic studies, preferably from more than one center or group II-3 Multiple time series with or without the intervention (also includes dramatic results in uncontrolled experiments) III Opinions of respected authorities, based on clinical experience, descriptive studies and case reports, or reports of expert committees • Aggregate internal validity is the degree to which the study(ies) provides valid evidence for the population and setting in which it was conducted • Aggregate external validity is the extent to which the evidence is relevant and generalizable to the population and conditions of typical primary care practice • Coherence/consistency	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
EPC Quality Assessments						
Chestnut et al., 1999[60]	Based on hierarchy of research design considered design and execution as well	Typically more than one	NA		Class I : Properly designed randomized controlled trials Class II: IIA Randomized controlled trials that contain design flaws preventing a specification of Class I IIA Multicenter or population-based longitudinal (cohort) studies IIB Controlled trials that were not randomized IIB Case-control studies IIB Case series with adequate description of the patient population, interventions, and outcomes measured. Class III: - Descriptive studies (uncontrolled case series) - Expert opinion - Case reports - Clinical experience	Grading is for quality of evidence and is based primarily on research design.

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
West et al., 1999[65] Pharmacological Treatment of Alcohol Dependence (RTI-UNC EPC)	Based on methodology, conduct, and analysis	Considers sample size and magnitude of difference in efficacy between intervention and placebo	Incorporates consistency among studies		Grades: A (good) Sufficient data for evaluating efficacy; sample size is adequate; data are consistent and indicate that the key drug is clearly superior to placebo for treatment of alcohol dependence. B (fair) Sufficient data for evaluating efficacy; sample size is adequate; data indicate inconsistencies in findings for alcohol outcomes between the drug and placebo such that efficacy of the key drug for treatment of alcohol dependence is not clearly established. C (poor) Sufficient and consistent evidence that the key drug is no more efficacious for treating alcohol dependence than placebo; sample size is adequate.	Note: Primarily concerns RCTs because only one non-RCT was included in the analysis
McNamara et al., 2001[66] Management of New Onset Atrial Fibrillation (JHU EPC)	NA	Strength of evidence depends on estimated magnitude of effect, precision of estimate, and confidence that there is a true effect	NA		System of grading dependent upon OR and CI: Evidence of efficacy: Strong OR>1.0, 99% CI does not include 1.0 Moderate OR>1.0, 95% CI does not include 1.0, but 99% CI does Suggestive 95% CI includes 1.0 in the lower tail (0.05<p<0.2-0.3) and the OR is in a clinically meaningful range Inconclusive 95% CI widely distributed around 1.0 Evidence of Lack of Efficacy: Strong OR near 1.0, 95% CI is narrow	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Ross et al., 2001[67] Management of Newly Diagnosed Patients with Epilepsy (Metaworks, Inc)	Based on hierarchy of research design	Number of studies and power of studies	NA		Levels of evidence: I Evidence obtained from meta-analysis of multiple, well-designed, controlled studies or from high-power RCTs II Evidence obtained from at least one well-designed experimental study or low power RCT III Evidence obtained from well-designed, quasi-experimental studies such as nonrandomized, controlled single group, pre-post, cohort, time, or matched case-control series IV Evidence from well-designed, nonexperimental studies, such as comparative and correlational descriptive and case studies V Evidence from case reports and clinical examples	Evidence scores for individual studies were computed by dividing the Jadad score by the level of evidence.

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Levine et al., 2000[147] Diagnosis and Management of Breast Disease (Metaworks, Inc)	Based on hierarchy of research design	Number of studies and power of studies	NA		I Evidence based on RCTs (or MA of RCT) of adequate size to ensure a low risk of incorporating false-positive or false-negative results II Evidence based on RCTs that are too small to provide level I evidence. These may show either positive trends that are not statistically significant or no trends and are associated with a high risk of false-negative results. III Evidence based on nonrandomized, controlled or cohort studies, case series, case-controlled studies or cross-sectional studies IV Evidence based on the opinion of respected authorities or that of expert committees as indicated in published consensus conferences or guidelines V Evidence which expresses the opinion of those individuals who have written and reviewed these guidelines, based on their experience, knowledge of the relevant literature and discussion with their peers	

Grid 5B. Overall Description of Systems to Grade Strength of Evidence (cont'd)

	Domain					
	Quality	**Quantity**	**Consistency**	**Other**	**Strength of Evidence Grading System**	**Comments**
Source						
Goudas et al, "Chapter 2. Methods." Management of Cancer Pain, 2000[68] and Lau et al., "Chapter 2. Methods." Evaluating Technologies for Identifying ACI in ED, 2000[59]	Internal validity graded on a 4 category system based on design and likelihood of bias (see details under system column)	Study size and magnitude of treatment effect	NA	Applicability of the evidence from study populations to the population at large	Internal validity of RCTs: A Double-blinded, well-concealed randomization, few drop outs, and no (or only minor reporting problem of the trial that is likely to cause significant bias B Single-blinded only, unclear concealment of randomization, or has some inconsistency in the reporting of the trial but is unlikely to result in major bias C Unblinded study, inadequate concealment of random allocation, high drop out rate, or has substantial inconsistencies in the reporting of the trial such that it may result in large bias D Inadequately reported (very often trials do not report certain data; this may occur by intent or due to oversight) Internal validity of non-randomized studies graded on study design and adequacy of reporting: A Prospective controlled trial B Cohort C Case-series	

Appendix D: An Annotated Bibliography of Empiric Evidence Used in Grid Development

We use the term "empiric evidence" in this report to mean aspects of study design, conduct, and analysis that have been shown, via methodological studies, to be related to risk of bias. When these aspects are not addressed or are poorly addressed in a study, it is more likely that the results from this study will give false or misleading results. For Tables 7-10 in this report (Chapter 2, Methods) we have designated particular domains and elements as empirically based. Exhibit D-1 (at the end of this appendix) catalogs the empirical evidence that we have used to arrive at these designations.

We acknowledge that there is disagreement between respected methodological experts, epidemiologists, and statisticians on some of these issues; we have attempted to take a moderate approach. Where empirical evidence was available but contradictory on a given domain or element topic, we elected not to define an empiric position on that topic. Where evidence was scant but clear, we included it as empiric but emphasize that future research may alter our conclusion.

A thorough assessment of underlying empiric evidence was not among the objectives of this project. Rather, this appendix arose from our need to categorize and make sense of the relevant research base. Although the information is fairly comprehensive, we have not undertaken the steps necessary to assure that it is exhaustive.

Systematic Reviews and Meta-Analyses

Literature Searches

Searches need to be comprehensive to assure that all relevant studies are included in a systematic review. Searches that rely on computerized databases such as Medline© are not likely to find all relevant studies.[149] Related issues are those of publication bias and country of origin of the study.

Publication Bias

Publication bias refers to the phenomenon that "positive studies" (e.g., studies that find a particular therapy works) are more likely to be published than "negative studies" (which do not find that the therapy is effective); unpublished studies are difficult to locate.[150-152] Studies funded by the pharmaceutical industry may be published less often than studies with other sources of funding—a type of publication bias.[151] Thus, a systematic review or meta-analysis of only the published studies may be misleading, producing a more favorable summary estimate than would have occurred if the entire body of literature was summarized, including published and unpublished works.

Language and Country of Origin

For a variety of reasons including cost and simplicity, many searches are often restricted to English language only. Moher and colleagues found no significant differences in completeness of reporting of key study elements for Randomized Controlled Trials (RCTs) published in English versus other languages.[153] Another study by Moher et al.[154] found no evidence that language-restricted meta-analyses were biased in terms of estimates of efficacy, but adding non-English RCTs did yield more precise estimates of effect.

For at least some types of studies, the results of the study reflect where the study was conducted. Vickers et al. found that trials of acupuncture from China, Japan, Hong Kong, Taiwan, and USSR/Russia were positive in all but one case.[155] Studies of interventions other than acupuncture originating from these countries were also overwhelmingly likely to find a positive effect of the intervention. Most experts believe that this pattern is a form of publication bias as discussed above. However, how a body of literature that contains studies from these countries should be handled in a systematic review is not clear. Our criterion in Table 7 specified that if investigators restrict their searches on the basis of language or country of origin, then they should provide some justification for this decision.

Masking (Blinding) of Reviewers

Evidence is conflicting about whether masking quality assessment reviewers to the authors of the study minimizes bias in a systematic review. Jadad et al. found that quality scores were lower and more consistent when reviewers were masked,[34] but Moher et al. found that quality scores were higher with masked quality assessment.[41] Two other methodological studies have found that quality scores did not differ significantly when reviewers were masked compared with open assessment. [95,156] A third study found no effect of reviewer masking on the summary measure of effect in meta-analysis.[157] Overall, we concluded that the evidence was insufficient to substantiate reviewer masking as a necessary and empirically supported quality element.

Quality Assessment

Some type of quality assessment of the individual studies that go into a systematic review is needed; however, the techniques for assessing study quality have not been well defined and there is conflicting evidence among the studies addressing this issue. Emerson and colleagues did not find that differences between treatments were related either to quality scores using the Chalmers scale or to results using an individual quality components approach.[158]

A study of quality assessment for RCTs comparing standard versus low molecular weight heparin (LMWH) to prevent post-operative thrombosis (DVT) by Juni and colleagues provided evidence that quality assessment scales weight components of quality differently.[2] They applied 25 different scales to each of the 17 RCTs in the meta-analysis and found that the summary relative risk for each scale differed, depending on whether high quality or low quality scales were evaluated. Whether LMWH was superior to regular heparin depended on which quality scale was used and the actual quality score. Using meta-regression techniques, they performed a

component-only analysis that focused on randomization, allocation concealment, and handling of withdrawals, showing that these quality components were not significantly associated with treatment effect. However, masking of outcome assessment is a critical quality component when comparing LMWH and regular heparin because tests to detect DVT are somewhat subjective.

Khan and colleagues reported that lower quality studies were more likely to find a positive effect of fertility treatment whereas higher quality studies did not.[35] An extensive methodological study by Moher et al. also found that meta-analyses using only low-quality RCTs had significantly higher effect estimates that meta-analyses using only high-quality studies.[41] Moher and colleagues found that, on average, low-quality RCTs found a 52% treatment benefit whereas high-quality studies found only a 29% benefit. Moher's study, which cuts across types of interventions and fields of medicine, offers the strongest evidence on this topic.

Although no one scale is likely to provide the best quality assessment in all cases, some aspects of study design, conduct, and analysis are related to study bias, and these quality items should be assessed as part of the process of conducting a systematic review or meta-analysis. However, we acknowledge that there is more empirical evidence supporting these quality components from the RCT literature, some of which was addressed in our discussion above and will be supported in the following section on empirical evidence relating to RCTs.

Heterogeneity

One reason that apparently similar studies do not find similar results is the degree of heterogeneity among them. Heterogeneity refers to differences in estimates of effect that are related to particular characteristics of the population or intervention studied. Thompson evaluated meta-analyses for cardiac and cancer outcomes and studies of cholesterol lowering.[159] He found that the conclusions of meta-analyses might differ if heterogeneity (due to such factors as age of study participants or duration of treatment) is not considered. This study supports what has long been considered "good practice" for systematic reviews, that a careful assessment of the similarities and differences among studies should be undertaken before studies are combined in a systematic review or meta-analysis. Statistical pooling of study results using meta-analytic techniques may not be advisable when substantial heterogeneity is present, but heterogeneity may provide important clues to explain treatment variation among subgroups of the population.[157]

Funding and Sponsorship

We found sufficient empirical evidence that funding and sponsorship of systematic reviews was related to the reporting of treatment effect. Barnes and Bero reported that systematic reviews of observational studies of the effects of passive tobacco smoke exposure were more likely *not* to find an adverse health effect if the authors had affiliations with the tobacco industry.[3] A similar study by Stelfox and colleagues found that authors with financial affiliations to the pharmaceutical industry were significantly more likely to endorse the safety of calcium channel blockers.[110] However, we do not support the view that the results of studies where authors received support from non-government sources are inherently biased. Rather, we believe that the

important principle is whether the authors of a study have competing interests sufficient to bias the results of the study—financial relationships are clearly only one such potential competing interest.

Randomized Controlled Trials

Randomization

A large and long-standing empirical body of evidence supports the superiority of RCTs for measuring treatment effect compared with nonrandomized designs. [27,105,160] As a study design element, randomization is powerful because it minimizes selection bias, thus increasing the likelihood that differences among treatment groups are actually the result of the treatment rather than some other prognostic factor.

The randomization domain seen on Table 8 and Grid 2 includes three empirically based elements: an adequate approach to sequence generation and appropriate allocation concealment, both of which result in group comparability at baseline. Studies of these three elements may overlap; some also address the issue of double- or triple-blinding. The process of randomization has two distinct parts. The first is how the random sequence is produced and the second is how patients' treatment group allocation is concealed. Methods of generating the sequence that are not truly random (e.g., using odd and even year of birth) and methods of concealment that can be subverted (e.g., peeking inside assignment envelopes) may allow investigators or clinicians to "rig" the study groups. This may result in study groups that are not similar in terms of their prognostic factors at baseline.

Schulz and colleagues reported that only one-third of RCTs in obstetrics and gynecology reported an adequate method of randomization.[161] They noted that observed differences in the baseline characteristics of study groups further suggested that randomization was improperly done. Studies that failed to report an adequate approach to sequence generation were unlikely to report adequate allocation concealment, and nearly half of the studies did not report an adequate method of allocation concealment.[162]

Allocation concealment may be more important than the exact procedures for generating the randomization sequence. Chalmers et al. found substantial case fatality differences among studies of treatments for myocardial infarction depending on whether the study was randomized and whether allocation was concealed.[105] Case fatality rate differences were 8.8% for studies that were randomized and properly concealed, 24.4% for unblinded randomized studies, and 58.1% for nonrandomized studies in cardiology, neurology, and pulmonology. Moher and colleagues found that trials with inadequately reported allocation concealment had significantly exaggerated estimates of treatment effect compared with studies that adequately reported concealment. [41]

Blinding

Allocation concealment inherently implies blinded assessment. Although usage differs, "single-blinding" generally refers to the study subject or patient not being aware of the treatment

allocation, whereas “double-blinding” typically means that neither the patient nor the caregivers know the treatment group assignment. However, the principle of double-blinding more generally means that the treatment assigned and received is masked to all key study personnel (e.g., investigators, caregivers, subjects, outcome assessors, data analysts) as well as participants. The study by Colditz et al. found that RCTs that did not employ double-blinding were significantly more likely to show a treatment effect.[27] Not all interventions can be successfully blinded; for health services research, it is difficult to mask participants and caregivers to factors such as their type of health care coverage or the type of clinician caring for them. Just as not all interventions can be randomized, not all interventions can be kept from those who are participating in the study.

Statistical Analysis

As in any study design, bias can be introduced at any point from design to reporting but the analysis strategy for RCTs is key. It is rare for studies to have totally complete follow-up of participants, and subjects leave the study for a variety of reasons. If the reason for a subject’s withdrawal is related to the therapy received or the outcome of interest, then bias may be introduced. If the study is analyzed on the basis of which treatment was actually received (an efficacy analysis) rather than by treatment assigned (an intent-to-treat analysis) then randomization is not maintained. Bias is even further increased when less adherent patients have significantly different outcomes and adherence is related to group assignment; underlying prognostic characteristics may be related to adherence and/or treatment effect, as well.

Chene and colleagues examined withdrawal issues, comparing an intent-to-treat analysis with an efficacy analysis in an HIV drug study. The relationship between adherence to the drug and outcomes was significant. The intent-to-treat analysis indicated that drug was not effective, which was not supported by the efficacy analysis.[163] Lachin reported similar results in a study of an Alzheimer’s drug where substantial numbers of participants withdrew from the RCT because of drug side effects.[164] Both the efficacy and intent-to-treat analyses supported the new drug, but only the latter supported its effectiveness at higher doses.

These statistical challenges are similar to those noted by Khan and colleagues comparing crossover trials to parallel-group RCTs evaluating infertility interventions.[35] They found that crossover trials overestimated effectiveness by an average of 74%—subjects who became pregnant were no longer eligible to be “crossed over” to the next treatment in the sequence of treatments being tested.

Funding and Sponsorship

RCTs may be subject to bias related to the author’s competing interests. Djulbegovic et al. found that pharmaceutical industry-sponsored studies were more likely to result in favorable evaluations of new treatments.[165] That studies conducted to support the efficacy of new treatments tend to show more favorable results is consistent with the drug approval process. Because of the expense, large phase III studies to support regulatory approval will only be conducted if the pharmaceutical company is relatively certain that its new treatment is

efficacious. However, this may not be the situation for smaller RCTs where not as much financial investment is involved; an example is the comparison between brand-name and generic levothyroxine for treating hypothyroidism.[166,167]

Djulbegovic and colleagues also noted that the choice of a comparative therapy known or suspected of being less effective—that is, in violation of the equipoise principle—might account for much of the bias found.[165] A study by Cho and Bero has been used to support the potential for conflict of interest based on funding sources. They found that studies published in pharmaceutical company-sponsored symposia proceedings were significantly more likely to favor the new drug of interest than were studies published in peer-reviewed journals.[168]

Observational Studies

As discussed in previous sections, empirical evidence clearly guides quality assessment of systematic reviews and RCTs. By contrast, little evidence helps guide the evaluation of observational studies beyond good epidemiologic practice and principles. Comparability of subjects was the only empirically derived element we designated for observational studies, relating to the use of concurrent versus historical controls groups. Chalmers et al. noted that the use of nonrandomized trials with historical controls exaggerated treatment effects in studies of anticoagulation for acute myocardial infarction.[160] Concato, Shah, and Horowitz compared RCTs and observational studies using concurrent control groups for five clinical topic areas (BCG vaccine for tuberculosis, mammography to prevent breast cancer deaths, cholesterol lowering and the risk of trauma mortality, hypertension treatment, and the risks of both stroke and coronary heart disease).[169] They found that estimates of effect were similar for RCTs and observational studies when the observational studies were rigorous i.e., using concurrent controls.

Two studies provide empirical evidence of bias in observational studies related to competing interests, which we have termed funding and sponsorship. The Cho and Bero study noted that both RCTs and observational studies reported in symposia proceedings tended to show favorable treatment effects.[168] In a similar study comparing the publications found in symposia proceedings versus peer-reviewed journals, articles in symposia were more likely to have been supported by the tobacco industry and less likely to have government funding.[170] Multivariate analysis indicated that peer-review was an important quality criterion rather than source of funding. This study lends support for a quality criterion of peer-review as an empirically based domain.

Diagnostic Studies

The domains and elements we used to compare tools to evaluate the quality of diagnostic studies were meant to be supplemental to those considered for RCTs and observational studies, as these are the two designs typically employed to evaluate diagnostic tests. The domains that we derived for diagnostic studies are unique; all have some empirical basis as a result of the work of Lijmer and colleagues.[78] They evaluated whether certain design factors perceived as "good practice" influenced the risk of bias. Of the five study design factors to be associated with bias, studies that evaluated the test in persons with known disease status showed more biased results

than if the test had been evaluated in a population with a full spectrum of disease. Studies that used a different reference standard for confirmation of positive and negative test results and those that interpreted the reference standard with full knowledge of the test result were also subject to substantial bias. The work of Lachs and colleagues supported that of Lijmer et al. in that the key test characteristics of sensitivity and specificity were affected by the spectrum of disease in the population tested.[171]

Exhibit D-1. Empirical Evidence Used to Derive Study Quality Domains

Source	Methodologic Issue Studied	Study Design Addressed	Summary of Findings
Chalmers et al., 1977[160]	RCTs vs. nonrandomized controlled trials using historical controls	Controlled trials	Use of historical controls in nonrandomized controlled trials of the use of anticoagulants for myocardial infarction led to exaggerated estimates of mortality reduction compared with RCT study designs.
Chalmers et al., 1983[105]	Randomization blinding (i.e., allocation concealment) in therapeutic trials of treatment for acute myocardial infarction	RCT	Case fatality differences were 8.8% in blinded randomization studies, 24.4% in unblinded randomized studies, and 58.1% in non-randomized studies. Evidence to support randomized study designs with double-blinding to minimize bias.
Simes, 1986[150]	Publication bias in clinical oncology	Systematic review*	Analysis of all published trials yielded increased estimates of effect for "new" therapies compared with analysis of trials registered in advance of conduct with an international registry.
Colditz et al., 1989[27]	Randomized versus non-randomized and double-blinded versus non-blinded trials in cardiology, neurology, respiratory medicine, and psychiatry.	RCT	Non-randomized sequential studies found larger therapeutic gains for the innovation compared to standard therapy ($p = 0.004$). RCTs that did not employ double-blinding had a higher likelihood of showing a positive effect of the innovation ($p = 0.02$). Evidence to support randomized study designs with double-blinding to minimize bias.
Emerson et al., 1990[158]	Relationship between study quality using the Chalmers scale and treatment differences in RCTs (primarily in various meta-analyses of cardiovascular trials, but with one dataset each of progesterone therapy in pregnancy, nicotine chewing gum for smoking cessation, and antibiotic therapy for GI surgery)	Systematic review*	No relationship between quality scores (using the entire scale) and treatment differences or variation in treatment difference was found. Using a component approach, inclusion of randomization blinding and/or handling of withdrawals was not associated with treatment differences either.

Exhibit D-1. Empirical Evidence Used to Derive Study Quality Domains (continued)

Source	Methodologic Issue Studied	Study Design Addressed	Summary of Findings
Easterbrook et al., 1991[151]	Publication bias	Systematic review	Study of research projects approved by a central ethics committee between 1984 and 1987 found that studies with significant results, non-randomized trials, observational studies, and laboratory-based trials were significantly more likely to be published. Studies funded by the pharmaceutical industry were less likely to be published than studies with other types of funding.
Lachs et al.[171]	Spectrum bias in diagnostic tests	Diagnostic tests	Sensitivity and specificity of urine dip stick for diagnosis of UTI differed markedly between groups of patients at high and low pre-test risk for UTI. The spectrum of disease in the patient population affects test characteristics and thus is important when evaluating a diagnostic test.
Dickersin et al., 1994[149]	Searching for RCTs in ophthalmology	Systematic review*	Medline® searches are not sufficiently sensitive to obtain all RCTs in field secondary to inadequate indexing, incomplete coverage of medical literature by Medline, skill level of searcher, and unpublished trials.
Thompson, 1994[159]	Heterogeneity in meta-analyses of cardiac, cancer outcomes, and cholesterol lowering	Systematic eview	Conclusions of meta-analyses may differ if heterogeneity among studies exists (due to issues such as age of subjects, duration of therapy, extent of cholesterol reduction, and confounding due to tobacco use).
Jeng et al., 1995[152]	Meta-analysis using individual patient data versus summary data from published and unpublished	Systematic review	The effect of treatment for infertility using paternal white blood cell immunization for recurrent miscarriage was statistically significant for pooled summary data from published studies with diminishing estimates of effect for meta-analysis using individual patient data or meta-analysis using unpublished data.
Cho and Bero, 1996[168]	Drug studies published in symposium proceedings	RCT, Observational	Studies sponsored by pharmaceutical companies and published in symposium proceedings were more likely to report favorable effects of the drug of interest than were studies published under peer review.
Schulz et al., 1994[161]	Randomization sequence generation, allocation concealment, and baseline characteristics in obstetrics and gynecology trials	RCT	Only about a third (32%) of trials reported an adequate method of sequence generation, and nearly half (48%) did not report methods used to conceal allocation. Only 9% reported adequate techniques for both. Differences in baseline characteristics among study groups in unrestricted trials were smaller than what would be statistically expected if randomization had been done properly.

Exhibit D-1. Empirical Evidence Used to Derive Study Quality Domains (continued)

Source	Methodologic Issue Studied	Study Design Addressed	Summary of Findings
Grimes and Schulz, 1996 [162]	Reporting of randomization sequence generation and allocation concealment for RCTs in obstetrics and gynecology	RCT	Failure to report an adequate approach to sequence generation was highly associated with failure to report adequate allocation concealment (p <0.001).
Jadad et al., 1996[34]	Need for blinded quality assessment of studies in systematic reviews. Quality assessment included items on randomization, double-blinding, and handling of withdrawals/dropouts	Systematic review*	Blind assessment resulted in lower and more consistent quality assessments.
Khan et al., 1996[35]	Crossover trials versus parallel group design in infertility research	RCT	Crossover trials overestimated odds ratios(ORs) by 74% (95% Confidence Interval [CI]: 2% to 197%) compared with parallel study designs evaluating the same interventions.
Khan et al., 1996[97]	Study quality and bias in systematic reviews of antiestrogen therapy for oliospermia	Systematic review	High quality studies did not find evidence of effectiveness, while low quality studies did. The overall summary OR for all studies had a positive OR, but a CI that crossed 1.
Moher et al., 1996[153]	Non-English language trials	Systematic review*	No significant differences for completeness of reporting of key study elements (randomization, double-blinding, withdrawals) for trials published in English versus other languages
Vickers et al., 1998[155]	Positive trial results and country of origin of study	Systematic review*	Trials of acupuncture originating in China, Japan, Hong Kong, Taiwan, and Russia/USSR had positive findings in all but one case. For trials of interventions other than acupuncture, publication of positive results occurred 99%, 89%, 97%, and 95% for studies originating in China, Japan, Russia/USSR, and Taiwan, respectively. No trial published in China or Russia/USSR found a treatment to be ineffective.

Exhibit D-1. Empirical Evidence Used to Derive Study Quality Domains (continued)

Source	Methodologic Issue Studied	Study Design Addressed	Summary of Findings
Barnes and Bero, 1997[170]	Quality of peer-reviewed original research publications versus non-peer-reviewed articles published in symposium proceedings Funding/Support	Primarily observational	Symposium articles on the health effects of environmental tobacco smoke exposure were found to be of poorer quality than peer-reviewed articles using a multivariate model which controlled for study design, article conclusion, article conclusion, article topic, and whether the source of funding was acknowledged. Symposium articles were significantly more likely to have tobacco industry funding or to have no source of funding acknowledged and less likely to have government funding. However, in multivariate modeling, funding source *per se* was not found to be significant.
Berlin, 1997[157]	Blinding of reviewers to journal, author, institution, and treatment group for meta-analysis of RCTs	Systematic review*	Blinding of reviewers during study selection and data extraction, using document scanning and editing, had neither a clinically nor a statistically significant effect on the summary odds ratios for meta-analyses of five different medical interventions.
Barnes and Bero, 1998[3]	Author affiliation and conclusions of reviews of effects of passive smoke exposure	Systematic review†	Reviews that found passive smoke exposure not to be associated with adverse health effects largely had authors with tobacco industry affiliation.
Chene et al., 1998[163]	Intention-to-treat (ITT) statistical analysis	RCT	A significant interaction between compliance and treatment outcome was found in this study of pyrimethamine prophylaxis of cerebral toxoplasmosis in HIV-infected patients. The ITT analysis did not show a significant treatment effect, while the on-treatment efficacy analysis did show a positive effect of the drug. The authors firmly believe that ITT analysis provides the only interpretable analysis of RCTs based on the following rationale: (1) randomization is maintained by an ITT analysis, (2) bias may result in an efficacy analysis when noncompliant patients have poorer outcomes and an interaction exists between compliance and treatment group, (3) prognostic factors affect compliance and treatment effect cannot be taken into account in an efficacy analysis, and (4) generalization is impossible without an ITT analysis.

Exhibit D-1. Empirical Evidence Used to Derive Study Quality Domains (continued)

Source	Methodologic Issue Studied	Study Design Addressed	Summary of Findings
Moher et al,1998[41]	Masked versus unmasked RCT study quality assessment	Systematic review*	Masked study quality assessment resulted in study quality scores were higher and statistically different (3.8% difference, p=0.005) compared with open assessment.
	Allocation concealment	RCT	Trials with inadequate reporting of allocation concealment had statistically exaggerated estimates of treatment effect, where the ratio of odds ratios was: 0.63, [95% CI 0.45, 0.88].
	Incorporation of study quality into meta-analyses	Systematic review*	Meta-analysis using only low quality trails had significantly greater estimate of treatment effect compared with meta-analysis of only high quality trials. Use of a quality weight in meta-regression rather than analyzing only low or high quality studies independently resulted in an estimate that had the least statistical heterogeneity and that was similar to the average treatment benefit of all trials, regardless of quality.
Stelfox et al., 1998[110]	Industry funding/sponsorship of research	Various	This study examined 5 original research articles, 32 review articles, and 33 letters to the editor published between March 1995 and September 1996 that had information about the safety of calcium-channel antagonists. 96% of authors supportive of calcium-channel antagonist safety had financial relationships with manufacturers compared with 60% of authors with neutral positions and 37% of authors who were critical of the safety of these agents (p <0.001). Supportive and neutral authors were also more likely than critical authors to have financial interactions with manufacturers of competing products. 100% of supportive, 67% of neutral, and 43% of critical authors had financial interactions with any pharmaceutical manufacturers (p <0.001).

Exhibit D-1. Empirical Evidence Used to Derive Study Quality Domains (continued)

Source	Methodologic Issue Studied	Study Design Addressed	Summary of Findings
Verhagen et al., 1998[156]	Blinding of balneotherapy study quality assessment using the Maastricht criteria	Systematic review*	Quality scores assessed using blinded versus nonblinded reviewers did not differ significantly.
Clark et al., 1999[95]	Reviewer blinding and use of the Jadad scale to rate the quality of studies on technologies to reduce perioperative allogenic blood transfusions	Systematic review*	Reviewer blinding did not result in a consistently significant effect on quality assessment. Found considerable interrater variability when using the Jadad scale, largely because of disagreements on the withdrawal item.
Juni et al., 1999[2]	Relationship of quality assessment using 25 different scales to treatment effects in a meta-analysis of 17 RCTs comparing standard to low molecular weight heparin (LMWH) for prevention of postoperative thrombosis	Systematic review*	6 scales found LMWH superior to standard heparin only in low quality trials; 7 scales found LMWH superior only in high quality trials; and the summary quality scores using the remaining 12 scales found similar estimates of effect in both high and low quality study strata. Using component approaches only found no significant association of treatment effect and allocation concealment or handling of withdrawals. However, open outcome assessment was associated with exaggerated treatment estimates (35% on average).
Lijmer et al., 1999[78]	Design of diagnostic test studies and risk of bias	Diagnostic tests	Evidence of exaggerated performance of diagnostic tests was found for studies with the following design flaws: 1. Evaluating the test in a diseased population and a separate control group (relative diagnostic odds ratios ([RDOR]: 3.0 [95% CI 2.0, 4.5]); 2. Use of a different reference standard for confirmation of positive and negative results of the test under study (RDOR: 2.2 [1.5, 3.3]); 3. Interpretation of the reference standard with knowledge of the test result (RDOR: 1.3 [1.0, 1.9]); 4. Lack of description of the test (RDOR: 1.7 [1.1, 2.5]); and 5. No description of the study population (RDOR: 1.4 [1.1, 1.7]).

Exhibit D-1. Empirical Evidence Used to Derive Study Quality Domains (continued)

Source	Methodologic Issue Studied	Study Design Addressed	Summary of Findings
Concato et al., 2000[169]	Comparison of RCTs and well-designed observational studies using concurrent controls for five clinical topics (BCG vaccine for TB, mammography and mortality from breast cancer, cholesterol levels and trauma mortality, hypertension treatment and stroke, hypertension treatment and coronary disease)	Observational ‡	Estimates of effect were similar for RCTs and observational studies that used concurrent controls for each of the five clinical areas studied. All measures of effect had overlapping 95% CIs. For these clinical topics and cohort studies using concurrent controls it appears that meta-analyses of these types of rigorous observational studies come to the same conclusion as meta-analyses of RCTs.
Djulbegovic et al., 2000[165]	Pharmaceutical company sponsorship of RCTs	RCT	Biases toward new treatments were found in for-profit pharmaceutical industry-sponsored research may be due to violations of principles of equipoise (e.g., choice of an inappropriate comparative control).
Lachin, 2000[164]	Intent-to-treat (ITT) versus efficacy statistical analysis	RCT	This article compared an intention-to-treat (ITT) analysis with an efficacy analysis for an Alzheimer's disease drug trial where there were substantial drop-outs due to hepatotoxicity of the drug. Complete follow-up was available for 92% of the participants. While both the ITT and the efficacy analyses supported drug efficacy, the ITT analysis supported efficacy only at higher doses. The efficacy analysis introduced selection bias based on tolerance of and compliance with the drug.
Moher et al., 2000[154]	Non-English language trials	Systematic Review*	No evidence was found that language-restricted meta-analyses lead to biased estimates of treatment efficacy in 79 meta-analyses covering a wide variety of disease areas. The average difference between meta-analyses including and excluding non-English trials was 2% (ratio of odds ratios: 0.98, 95% CI 0.81, 1.17). Sensitivity analyses indicated that these findings were robust. Inclusion of non-English trials did result in more precise estimates of treatment efficacy, with CI averaging 16% narrower.

*Applies to systematic reviews of trials

†Applies to observational studies that are prospective cohort studies that use concurrent controls

‡Applies to systematic reviews of observational studies

Note: For complete reference information, see reference list

Appendix E: Excluded Articles

ID	Citation	Exclusion reason *[*]
3	"Assendelft, W. J.; Koes, B. W.; Knipschild, P. G., and Bouter, L. M. The relationship between methodological quality and conclusions in reviews of spinal manipulation. JAMA. 1995 Dec 27; 274(24):1942-8"	NR-Not able to abstract
7	"Barratt, A.; Irwig, L.; Glasziou, P.; Cumming, R. G.; Raffle, A.; Hicks, N.; Gray, J. A., and Guyatt, G. H. Users' guides to the medical literature: XVII. How to use guidelines and recommendations about screening. Evidence-Based Medicine Working Group. JAMA. 1999 Jun 2; 281(21):2029-34"	NR-Review
9	"Bastian, H. Raising the standard: practice guidelines and consumer participation. Int J Qual Health Care. 1996 Oct; 8(5):485-90"	NR-OCD
17	"Berlin, J. A., and Colditz, G. A. A meta-analysis of physical activity in the prevention of coronary heart disease. Am J Epidemiol. 1990 Oct; 132(4):612-28."	NR-ROS
23	"Bero, L. A.; Grilli, R.; Grimshaw, J. M.; Harvey, E.; Oxman, A. D., and Thomson, M. A. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. BMJ. 1998 Aug 15; 317(7156):465-8"	NR-ROS
25	"Briggs, A. H., and Gray, A. M. Handling uncertainty in economic evaluations of healthcare interventions. BMJ. 1999 Sep 4; 319(7210):635-8"	NR-ROS
27	"Bucher, H. C.; Guyatt, G. H.; Cook, D. J.; Holbrook, A., and McAlister, F. A. Users' guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. Evidence-Based Medicine Working Group. JAMA. 1999 Aug 25; 282(8):771-8"	NR-Implementation/Application
29	"Chalmers, I.; Adams, M.; Dickersin, K.; Hetherington, J.; Tarnow-Mordi, W.; Meinert, C.; Tonascia, S., and Chalmers, T. C. A cohort study of summary reports of controlled trials. JAMA. 1990 Mar 9; 263(10):1401-5"	NR-ROS
39	"Clarke, M. The QUORUM statement. Lancet. 2000 Feb 26; 355(9205):756-7"	ECL
41	"Cluzeau, F.; Littlejohns, P.; Grimshaw, J., and Feder, G. Appraisal Instrument for Clinical Guidelines. St. George's Hospital Medical School; 1997 May"	NR-Not able to abstract

[*] See Table 4 for the code to abbreviations.

Excluded Articles (cont'd)

ID	Citation	Exclusion reason *[*]
49	"Cook, D. J.; Guyatt, G. H.; Ryan, G.; Clifton, J.; Buckingham, L.; Willan, A.; McIlroy, W., and Oxman, A. D. Should unpublished data be included in meta-analyses? Current convictions and controversies. JAMA. 1993 Jun 2; 269(21):2749-53"	NR-ROS
55	"Dans, A. L.; Dans, L. F.; Guyatt, G. H., and Richardson, S. Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. Evidence-Based Medicine Working Group. JAMA. 1998 Feb 18; 279(7):545-9"	NR-Implementation/Application
59	"Dickersin, K.; Higgins, K., and Meinert, C. L. Identification of meta-analyses. The need for standard terminology. Control Clin Trials. 1990 Feb; 11(1):52-66"	NR-ROS
73	"Fleiss, J. L., and Gross, A. J. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. J Clin Epidemiol. 1991; 44(2):127-39"	NR-Not able to abstract
75	"Garber, A. M. Realistic rigor in cost-effectiveness methods. Medical Decision Making. 1999 Oct-1999 Dec 31; 19(4):378-9; discussion 383-4"	ECL
79	"Giacomini, M. K., and Cook, D. J. Users' guides to the medical literature: XXIII. Qualitative research in health care B. What are the results and how do they help me care for my patients? Evidence-Based Medicine Working Group. JAMA. 2000 Jul 26; 284(4):478-82."	NR-Implementation/Application
89	"Guyatt, G. H.; Naylor, C. D.; Juniper, E.; Heyland, D. K.; Jaeschke, R., and Cook, D. J. Users' guides to the medical literature: XII. How to use articles about health-related quality of life. Evidence-Based Medicine Working Group. JAMA. 1997 Apr 16; 277(15):1232-7"	NR-Implementation/Application
91	"Guyatt, G. H., and Rennie, D. Users' guides to the medical literature. JAMA. 1993 Nov 3; 270(17):2096-7"	ECL
99	"Guyatt, G. H.; Sinclair, J.; Cook, D. J., and Glasziou, P. Users' guides to the medical literature: XVI. How to use a treatment recommendation. Evidence-Based Medicine Working Group and the Cochrane Applicability Methods Working Group. JAMA. 1999 May 19; 281(19):1836-43"	NR-Implementation/Application
103	"Harper, G.; Townsend, J., and Buxton, M. The preliminary economic evaluation of health technologies for the prioritization of health technology assessments. A discussion. International Journal of Technology Assessment in Health Care. 1998 Fall; 14(4):652-62"	NR-OCD
105	"Hayward, R. S.; Wilson, M. C.; Tunis, S. R.; Bass, E. B., and Guyatt, G. Users' guides to the medical literature: VIII. How to use clinical practice guidelines. A. Are the recommendations valid? The Evidence- Based Medicine Working Group. JAMA. 1995 Aug 16; 274(7):570-4"	NR-Implementation/Application
109	"Hill, S. R.; Mitchell, A. S., and Henry, D. A. Problems with the interpretation of pharmacoeconomic analyses: a review of submissions to the Australian Pharmaceutical Benefits Scheme. JAMA. 2000 Apr 26; 283(16):2116-21"	NR-ROS
111	"Hunt, D. L.; Jaeschke, R., and McKibbon, K. A. Users' guides to the medical literature: XXI. Using electronic health information resources in evidence-based practice. Evidence-Based Medicine Working Group. JAMA. 2000 Apr 12; 283(14):1875-9"	NR-Implementation/Application

[*] See Table 4 for the code to abbreviations.

Excluded Articles (cont'd)

ID	Citation	Exclusion reason *[*]
115	"Ioannidis, J. P., and Lau, J. Can quality of clinical trials and meta-analyses be quantified? Lancet. 1998 Aug 22; 352(9128):590-1"	ECL
121	"Jadad, A. R.; Cook, D. J.; Jones, A.; Klassen, T. P.; Tugwell, P.; Moher, M., and Moher, D. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. JAMA. 1998 Jul 15; 280(3):278-80"	NR-ROS
125	"Jaeschke, R.; Guyatt, G. H., and Sackett, D. L. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA. 1994 Mar 2; 271(9):703-7"	NR-Implementation/Application
127	"Kerridge, I.; Lowe, M., and Henry, D. Ethics and evidence based medicine. BMJ. 1998 Apr 11; 316(7138):1151-3."	NR-OCD
131	"Klassen, T. P.; Jadad, A. R., and Moher, D. Guides for reading and interpreting systematic reviews: I. Getting started. Arch Pediatr Adolesc Med. 1998 Jul; 152(7):700-4"	NR-OCD
137	"L'Abbe, K. A.; Detsky, A. S., and O'Rourke, K. Meta-analysis in clinical research. Ann Intern Med. 1987 Aug; 107(2):224-33"	NR-Not able to abstract
145	"Longnecker, M. P.; Berlin, J. A.; Orza, M. J., and Chalmers, T. C. A meta-analysis of alcohol consumption in relation to risk of breast cancer. JAMA. 1988 Aug 5; 260(5):652-6"	NR-Not able to abstract
147	"Mandelblatt, J. S.; Fryback, D. G.; Weinstein, M. C.; Russell, L. B., and Gold, M. R. Assessing the effectiveness of health interventions for cost-effectiveness analysis. Panel on Cost-Effectiveness in Health and Medicine. Journal of General Internal Medicine. 1997 Sep; 12(9):551-8."	NR-Review
149	"McAlister, F. A.; Laupacis, A.; Wells, G. A., and Sackett, D. L. Users' Guides to the Medical Literature: XIX. Applying clinical trial results. B. Guidelines for determining whether a drug is exerting (more than) a class effect. JAMA. 1999 Oct 13; 282(14):1371-7"	NR-Implementation/Application
151	"McAlister, F. A.; Straus, S. E.; Guyatt, G. H., and Haynes, R. B. Users' guides to the medical literature: XX. Integrating research evidence with the care of the individual patient. Evidence-Based Medicine Working Group. JAMA. 2000 Jun 7; 283(21):2829-36"	NR-Implementation/Application
153	"McGinn, T. G.; Guyatt, G. H.; Wyer, P. C.; Naylor, C. D.; Stiell, I. G., and Richardson, W. S. Users' guides to the medical literature: XXII. How to use articles about clinical decision rules. Evidence-Based Medicine Working Group. JAMA. 2000 Jul 5; 284(1):79-84"	NR-Not able to abstract
155	"Meltzer, D., and Johannesson, M. Inconsistencies in the 'societal perspective' on costs of the Panel on Cost-Effectiveness in Health and Medicine. Medical Decision Making. 1999 Oct-1999 Dec 31; 19(4):371-7."	NR-OCD
157	"Meltzer, D., and Johannesson, M. On the Role of Theory in Cost-Effectiveness Analysis-A Response to Garber, Russell, and Weinstein. Medical Decision Making. 1999 Oct-1999 Dec 31; 19(4):383-4."	ECL
161	"Milne, R., and Oliver, S. Evidence-based consumer health information: developing teaching in critical appraisal skills. Int J Qual Health Care. 1996 Oct; 8(5):439-45."	NR-OCD

[*] See Table 4 for the code to abbreviations.

Excluded Articles (cont'd)

ID	Citation	Exclusion reason *[*]
175	"Murphy, M. K.; Black, N. A.; Lamping, D. L.; McKee, C. M.; Sanderson, C. F. B.; Askham, J., and Marteau, T. Consensus development methods, and their use in clinical guideline development. Health Technology Assessment; 1998; pp. 55-61"	NR-Review
177	"Naylor, C. D., and Guyatt, G. H. Users' guides to the medical literature: X. How to use an article reporting variations in the outcomes of health services. The Evidence- Based Medicine Working Group. JAMA. 1996 Feb 21; 275(7):554-8"	NR-Implementation/Application
179	"Naylor, C. D., and Guyatt G. H. Users' guides to the medical literature: XI. How to use an article about a clinical utilization review. Evidence-Based Medicine Working Group. JAMA. 1996 May 8; 275(18):1435-9."	NR-Implementation/Application
181	"Nylenna, M. Details of patients' consent in studies should be reported. BMJ. 1997 Apr 12; 314(7087):1127-8"	ECL
183	"O'Brien, B. J.; Heyland, D.; Richardson, W. S.; Levine, M., and Drummond, M. F. Users' guides to the medical literature: XIII. How to use an article on economic analysis of clinical practice. B. What are the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA. 1997 Jun 11; 277(22):1802-6"	NR-Implementation/Application
197	"Oxman, A. D.; Sackett, D. L., and Guyatt, G. H. Users' guides to the medical literature: I. How to get started. Evidence-Based Medicine Working Group. JAMA. 1993 Nov 3; 270(17):2093-5"	NR-Implementation/Application
205	"Randolph, A. G.; Haynes, R. B.; Wyatt, J. C.; Cook, D. J., and Guyatt, G. H. Users' guides to the medical literature: XVIII. How to use an article evaluating the clinical impact of a computer-based clinical decision support system. JAMA. 1999 Jul 7; 282(1):67-74."	NR-Implementation/Application
209	"Rennie, D., and Luft, H. S. Pharmacoeconomic analyses: making them transparent, making them credible. JAMA. 2000 Apr 26; 283(16):2158-60."	ECL
211	"Richardson, W. S., and Detsky, A. S. Users' guides to the medical literature: VII. How to use a clinical decision analysis. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1995 Apr 26; 273(16):1292-5"	NR-Implementation/Application
213	"Richardson, W. S., and Detsky, A. S. Users' guides to the medical literature: VII. How to use a clinical decision analysis. B. What are the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA. 1995 May 24-1995 May 31; 273(20):1610-3"	NR-Implementation/Application
215	"Richardson, W. S.; Wilson, M. C.; Guyatt, G. H.; Cook, D. J., and Nishikawa, J. Users' guides to the medical literature: XV. How to use an article about disease probability for differential diagnosis. Evidence-Based Medicine Working Group. JAMA. 1999 Apr 7; 281(13):1214-9"	NR-Implementation/Application
217	"Richardson, W. S.; Wilson, M. C.; Williams, J. W. Jr.; Moyer, V. A., and Naylor, C. D. Users' guides to the medical literature: XXIV. How to use an article on the clinical manifestations of disease. Evidence-Based Medicine Working Group. JAMA. 2000 Aug 16; 284(7):869-75"	NR-Implementation/Application

[*] See Table 4 for the code to abbreviations.

Excluded Articles (cont'd)

ID	Citation	Exclusion reason *[*]
219	"Sacks, H. S.; Berrier, J.; Reitman, D.; Ancona-Berk, V. A., and Chalmers, T. C. Meta-analyses of randomized controlled trials. N Engl J Med. 1987 Feb 19; 316(8):450-5"	"NR-Other, newer version available"
223	"Sauerland, S., and Lefering, R. Quality of reports of randomised trials and estimates of treatment efficacy. Lancet. 1998 Nov 7; 352(9139):1555-6"	ECL
233	"Silagy, C. A. An analysis of review articles published in primary care journals. Fam Pract. 1993 Sep; 10(3):337-41"	NR-ROS
241	"Taddio, A.; Pain, T.; Fassos, F. F.; Boon, H.; Ilersich, A. L., and Einarson, T. R. Quality of nonstructured and structured abstracts of original research articles in the British Medical Journal, the Canadian Medical Association Journal and the Journal of the American Medical Association. CMAJ. 1994 May 15; 150(10):1611-5"	NR-ROS
245	"The Canadian Cooperative Study Group. A randomized trial of aspirin and sulfinpyrazone in threatened stroke. New Eng J Med. 1978 Jul 13; 299(2):53-9"	NR-ROS
259	"Whitman, N. I. The Delphi technique as an alternative for committee meetings. J Nurs Educ. 1990 Oct; 29(8):377-9"	NR-OCD
265	"Yusuf, S.; Peto, R.; Lewis, J.; Collins, R., and Sleight, P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Prog Cardiovasc Dis. 1985 Mar-1985 Apr 30; 27(5):335-71"	NR-Review
275	"Roman, S. H.; Silberzweig, S. B., and Siu, A. L. Grading the evidence for diabetes performance measures. Eff Clin Pract. 2000 Mar-2000 Apr 30; 3(2):85-91"	NR-Not able to abstract
279	"Woloshin, S. Arguing about grades. Eff Clin Pract. 2000 Mar-2000 Apr 30; 3(2):94-5"	ECL
295	"Lau, J.; Zucker, D.; Engles, E. A.; Balk, E.; Barza, M.; Terrin, N.; Devine, D.; Chew, P.; Lang, T. A., and Liu, D. Diagnosis and Treatment of Acute Bacterial Rhinosinusitis. Evidence Report/Technology Assessment No. 9. Agency for Health Care Policy and Research; 1999 Mar; AHCPR Publication No. 99-E016"	NR-Not able to abstract
301	"Diagnosis and Treatment of Swallowing Disorders (Dysphagia) in Acute-Care Stroke Patients. Evidence Report/Technology Assessment No. 8. Agency for Health Care Policy and Research; 1999 Jul; AHCPR Publication No. 99-E024"	NR-Not able to abstract
329	"How to read clinical journals: III. To learn the clinical course and prognosis of disease. Can Med Assoc J. 1981 Apr 1; 124(7):869-72"	NR-Implementation/Application
339	"Begg, C. B. Methodologic standards for diagnostic test assessment studies. J Gen Intern Med. 1988 Sep-1988 Oct 31; 3(5):518-20"	ECL
343	"Chalmers I. 'Applying overviews and meta-analyses at the bedside': Discussion. J Clin Epidemiol. 1995; 48(1):67-70"	NR-Other
347	"Evans, D. P.; Burke, M. S., and Newcombe, R. G. Medicines of choice in low back pain. Curr Med Res Opin. 1980; 6(8):540-7"	NR-ROS
349	"Faas, A.; Chavannes, A. W.; van Eijk, J. T., and Gubbels, J. W. A randomized, placebo-controlled trial of exercise therapy in patients with acute low back pain. Spine. 1993 Sep 1; 18(11):1388-95"	NR-ROS

[*] See Table 4 for the code to abbreviations.

Excluded Articles (cont'd)

ID	Citation	Exclusion reason *[*]
351	"Faas, A.; van Eijk, J. T.; Chavannes, A. W., and Gubbels, J. W. A randomized trial of exercise therapy in patients with acute low back pain. Efficacy on sickness absence. Spine. 1995 Apr 15; 20(8):941-7"	NR-ROS
353	"Farrell, J. P., and Twomey, L. T. Acute low back pain. Comparison of two conservative treatment approaches. Med J Aust. 1982 Feb 20; 1(4):160-4."	NR-ROS
363	"Hurlbut, T. A., 3d, and Littenberg, B. The diagnostic accuracy of rapid dipstick tests to predict urinary tract infection. Am J Clin Pathol. 1991 Nov; 96(5):582-8"	NR-ROS
369	"Littenberg, B., and Moses, L. E. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making. 1993 Oct-1993 Dec 31; 13(4):313-21"	NR-Stat Meth
373	"Moses, L. E.; Shapiro, D., and Littenberg, B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med. 1993 Jul 30; 12(14):1293-316"	NR-Stat Meth
385	"Toman, C.; Harrison, M.B., and Logan, J. Clinical practice guidelines: necessary but not sufficient for evidence-based patient education and counseling. Patient Education and Counseling. 2001 Mar; 42(3):279-87"	NR-OCD
393	"Wasson, J. H.; Sox, H. C.; Neff, R. K., and Goldman, L. Clinical prediction rules. Applications and methodological standards. N Engl J Med. 1985 Sep 26; 313(13):793-9"	NR-Not able to abstract
435	"Fishbain, D.; Cutler, R. B.; Rosomoff, H. L., and Rosomoff, R. S. What is the quality of the implemented meta-analytic procedures in chronic pain treatment meta-analyses? Clin J Pain. 2000 Mar; 16(1):73-85"	NR-ROS
447	"Johanson, R., and Lucking, L. Evidence-based medicine in obstetrics. Int J Gynaecol Obstet. 2001 Feb; 72(2):179-185"	NR-ROS
1008	"Beral, V. 'The practice of meta-analysis': discussion. Meta-analysis of observational studies: a case study of work in progress. J Clin Epidemiol. 1995 Jan; 48(1):165-6"	NR-OCD
1010	"Berkey, C. S.; Anderson, J. J., and Hoaglin, D. C. Multiple-outcome meta-analysis of clinical trials. Statistics in Medicine. 1996 Mar 15; 15(5):537-57"	NR-Stat Meth
1012	"Berkey, C. S.; Hoaglin, D. C.; Antczak-Bouckoms, A.; Mosteller, F., and Colditz, G. A. Meta-analysis of multiple outcomes by regression with random effects. Statistics in Medicine. 1998 Nov 30; 17(22):2537-50"	NR-Stat Meth
1014	"Berkey, C. S.; Hoaglin, D. C.; Mosteller, F., and Colditz, G. A. A random-effects regression model for meta-analysis. Statistics in Medicine. 1995 Feb 28; 14(4):395-411"	NR-Stat Meth
1016	"Boissel, J. and Cucherat, M. The meta-analysis of diagnostic test studies. European Radiology. 1998; 8(3):484-7"	NR-Review
1018	"Bramwell, V. H., and Williams, C. J. Do authors of review articles use systematic methods to identify, assess and synthesize information?. Annals of Oncology. 1997 Dec; 8(12):1185-95"	NR-ROS
1026	"Dean, M. Out of step with the Lancet homeopathy meta-analysis: more objections than objectivity?. Journal of Alternative & Complementary Medicine. 1998 Winter; 4(4):389-98"	NR-ROS

[*] See Table 4 for the code to abbreviations.

Excluded Articles (cont'd)

ID	Citation	Exclusion reason *[*]
1028	"Devine, E. C. Issues and challenges in coding interventions for meta-analysis of prevention research. NIDA Research Monograph. 1997; 170130-46"	NR-Design/Methods
1030	"Diezel, K.; Pharoah, F. M., and Adams, C. E. Abstracts of trials presented at the Vth World Congress of Psychiatry (Mexico, 1971): a cohort study. Psychological Medicine. 1999 Mar; 29(2):491-4"	NR-ROS
1040	"Gelskey, S. C. Cigarette smoking and periodontitis: methodology to assess the strength of evidence in support of a causal association. Community Dentistry & Oral Epidemiology. 1999 Feb; 27(1):16-24"	NR-Review
1044	"Hansen, W. B., and Rose, L. A. Issues in classification in meta-analysis in substance abuse prevention research. NIDA Research Monograph. 1997; 170183-201"	NR-ROS
1046	"Jadad, A. R.; Moher, D., and Klassen, T. P. Guides for reading and interpreting systematic reviews: II. How did the authors find the studies and assess their quality? Archives of Pediatrics & Adolescent Medicine. 1998 Aug; 152(8):812-7"	NR-Review
1048	"Jadad, A. R.; Moher, M.; Browman, G. P.; Booker, L.; Sigouin, C.; Fuentes, M., and Stevens, R. Systematic reviews and meta-analyses on treatment of asthma: critical evaluation. BMJ. 2000 Feb 26; 320(7234):537-40"	
1054	"Johansen, H. K., and Gotzsche, P. C. Problems in the design and reporting of trials of antifungal agents encountered during meta-analysis. JAMA. 1999 Nov 10; 282(18):1752-9"	NR-ROS
1056	"Jones, J. L. Drugs for AIDS/HIV: assessing the evidence. International Journal of Technology Assessment in Health Care. 1998 Summer; 14(3):567-72"	NR-Design/Methods
1060	"Kaegi, L. AMA Clinical Quality Improvement Forum ties it all together: from guidelines to measurement to analysis and back to guidelines. Joint Commission Journal on Quality Improvement. 1999 Feb; 25(2):95-106"	NR-Review
1062	"Kelly, S.; Berry, E.; Roderick, P.; Harris, K. M.; Cullingworth, J.; Gathercole, L.; Hutton, J., and Smith, M. A. The identification of bias in studies of the diagnostic performance of imaging modalities. British Journal of Radiology. 1997 Oct; 70(838):1028-35."	NR-Review
1066	"Ladhani, S. and Williams, H. C. The management of established postherpetic neuralgia: a comparison of the quality and content of traditional vs. systematic reviews. British Journal of Dermatology. 1998 Jul; 139(1):66-72."	NR-ROS
1068	"Lafata, J. E.; Koch, G. G., and Ward, R. E. Synthesizing evidence from multiple studies. The role of meta-analysis in pharmacoeconomics. Medical Care. 1996 Dec; 34(12 Suppl):DS136-45."	NR-Design/Methods
1070	"Lau, J.; Ioannidis, J. P., and Schmid, C. H. Quantitative synthesis in systematic reviews. Annals of Internal Medicine. 1997 Nov 1; 127(9):820-6."	NR-Design/Methods
1072	"Macarthur, C.; Foran, P. J., and Bailar, J. C. 3rd. Qualitative assessment of studies included in a meta-analysis: DES and the risk of pregnancy loss. Journal of Clinical Epidemiology. 1995 Jun; 48(6):739-47."	NR-ROS

[*] See Table 4 for the code to abbreviations.

Excluded Articles (cont'd)

ID	Citation	Exclusion reason *[*]
1074	"Meade, M. O. and Richardson, W. S. Selecting and appraising studies for a systematic review. Annals of Internal Medicine. 1997 Oct 1; 127(7):531-7."	NR-Review
1080	"Mulrow, C.; Langhorne, P., and Grimshaw, J. Integrating heterogeneous pieces of evidence in systematic reviews. Annals of Internal Medicine. 1997 Dec 1; 127(11):989-95"	NR-Not able to abstract
1082	"Myers, J. E., and Thompson, M. L. Meta-analysis and occupational epidemiology. Occupational Medicine. 1998 Feb; 48(2):99-101"	NR-Design/Methods
1084	"Olkin, I. Diagnostic statistical procedures in medical meta-analyses. Statistics in Medicine. 1999 Sep 15-1999 Sep 30; 18(17-18):2331-41"	NR-Stat Meth
1088	"Ramirez, A. J.; Westcombe, A. M.; Burgess, C. C.; Sutton, S.; Littlejohns, P., and Richards, M. A. Factors predicting delayed presentation of symptomatic breast cancer: a systematic review [see comments]. [Review] [34 refs]. Lancet. 1999 Apr 3; 353(9159):1127-31"	NR-ROS
1110	"Watt, D.; Verma, S., and Flynn, L. Wellness programs: a review of the evidence [see comments]. [Review] [34 refs]. Canadian Medical Association Journal. 1998 Jan 27; 158(2):224-30"	NR-ROS
2000	"British Association of Surgical Oncology Guidelines. The management of metastatic bone disease in the United Kingdom. Breast Specialty Group of the British Association of Surgical Oncology. Eur J Surg Oncol. 1999 Feb; 25(1):3-23"	NR-Guideline
2002	"Clinical practice guideline: diagnosis and evaluation of the child with attention-deficit/hyperactivity disorder. American Academy of Pediatrics. Pediatrics. 2000 May; 105(5):1158-70	NR-Guideline
2006	"Heart failure clinical guideline. South African Medical Association Heart Failure Working Group. S Afr Med J. 1998 Sep; 88(9 Pt 2):1133-55"	NR-Guideline
2008	"The management of minor closed head injury in children. Committee on Quality Improvement, American Academy of Pediatrics. Commission on Clinical Policies and Research, American Academy of Family Physicians. Pediatrics. 1999 Dec; 104(6):1407-15"	NR-Guideline
2010	"National Institutes of Health Consensus Development Conference Statement: Breast Cancer Screening for Women Ages 40-49, January 21-23, 1997. National Institutes of Health Consensus Development Panel. J Natl Cancer Inst. 1997 Jul 16; 89(14):1015-26"	NR-Guideline
2012	"Practice parameter: the diagnosis, treatment, and evaluation of the initial urinary tract infection in febrile infants and young children. American Academy of Pediatrics. Committee on Quality Improvement. Subcommittee on Urinary Tract Infection. Pediatrics. 1999 Apr; 103(4 Pt 1):843-52"	NR-Guideline
2014	"Practice parameter: the management of acute gastroenteritis in young children. American Academy of Pediatrics, Provisional Committee on Quality Improvement, Subcommittee on Acute Gastroenteritis. Pediatrics. 1996 Mar; 97(3):424-35"	NR-Guideline

[*] See Table 4 for the code to abbreviations.

Excluded Articles (cont'd)

ID	Citation	Exclusion reason *[*]
2016	"Recommendations for prevention and control of hepatitis C virus (HCV) infection and HCV-related chronic disease. Centers for Disease Control and Prevention. MMWR Morb Mortal Wkly Rep. 1998 Oct 16; 47(RR-19):1-39"	NR-Guideline
2018	"Vaccine-preventable diseases: improving vaccination coverage in children, adolescents, and adults. A report on recommendations from the Task Force on Community Preventive Services. MMWR Morb Mortal Wkly Rep. 1999 Jun 18; 48(RR-8):1-15"	NR-Guideline
2020	"Adams, J. L.; Fitzmaurice, D. A.; Heath, C. M.; Loudon, R. F.; Riaz, A.; Sterne, A., and Thomas, C. P. A novel method of guideline development for the diagnosis and management of mild to moderate hypertension. Br J Gen Pract. 1999 Mar; 49(440):175-9"	NR-Design/Methods
2022	"Anderson, I. M.; Nutt, D. J., and Deakin, J. F. Evidence-based guidelines for treating depressive disorders with antidepressants: a revision of the 1993 British Association for Psychopharmacology guidelines. British Association for Psychopharmacology. Journal of Psychopharmacology. 2000 Mar; 14(1):3-20"	NR-Guideline
2024	"Anderson, J. D. Need for evidence-based practice in prosthodontics. Journal of Prosthetic Dentistry. 2000 Jan; 83(1):58-65"	NR-Review
2030	"Begg, C. B. The role of meta-analysis in monitoring clinical trials. Statistics in Medicine. 1996 Jun 30; 15(12):1299-306; discussion 1307-11."	NR-OCD
2032	"Bernstein, S. J.; Hofer, T. P.; Meijler, A. P., and Rigter, H. Setting standards for effectiveness: a comparison of expert panels and decision analysis. International Journal for Quality in Health Care. 1997 Aug; 9(4):255-63"	NR-ROS
2036	"Bigby, M. Evidence-based medicine in a nutshell. A guide to finding and using the best evidence in caring for patients. Archives of Dermatology. 1998 Dec; 134(12):1609-18"	NR-Review
2038	"Bisno, A. L.; Gerber, M. A.; Gwaltney, J. M. Jr.; Kaplan, E. L., and Schwartz, R. H. Diagnosis and management of group A streptococcal pharyngitis: a practice guideline. Infectious Diseases Society of America. Clinical Infectious Diseases. 1997 Sep; 25(3):574-83"	NR-Guideline
2040	"Black, H. R., and Crocitto, M. T. Number needed to treat: solid science or a path to pernicious rationing? American Journal of Hypertension. 1998 Aug; 11(8 Pt 2):128S-134S; discussion 135S-137S"	NR-OCD
2042	"Black, W. C.; Nease, R. F. Jr., and Tosteson, A. N. Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. Journal of the National Cancer Institute. 1995 May 17; 87(10):720-31"	NR-ROS
2044	"Blumberg, J. B. Considerations of the scientific substantiation for antioxidant vitamins and beta-carotene in disease prevention. American Journal of Clinical Nutrition. 1995 Dec; 62(6 Suppl):1521S-1526S"	NR-Review
2050	"Burnand, B.; Vader, J. P.; Froehlich, F.; Dupriez, K.; Larequi-Lauber, T.; Pache, I.; Dubois, R. W.; Brook, R. H., and Gonvers, J. J. Reliability of panel-based guidelines for colonoscopy: an international comparison [see comments]. Gastrointestinal Endoscopy. 1998 Feb; 47(2):162-6"	NR-ROS

[*] See Table 4 for the code to abbreviations.

Excluded Articles (cont'd)

ID	Citation	Exclusion reason *[*]
2062	"Frank, C. Dementia workup. Deciding on laboratory testing for the elderly. Canadian Family Physician. 1998 Jul; 44:1489-95"	NR-Guideline
2064	"Freemantle, N.; Mason, J., and Eccles, M. Deriving treatment recommendations from evidence within randomized trials. The role and limitation of meta-analysis. International Journal of Technology Assessment in Health Care. 1999 Spring; 15(2):304-15"	NR-Design/Methods
2068	"Gershon, A. A.; Gardner, P.; Peter, G.; Nichols, K., and Orenstein, W. Quality standards for immunization. Guidelines from the Infectious Diseases Society of America. Clinical Infectious Diseases. 1997 Oct; 25(4):782-6"	NR-Guideline
2070	"Gibbons, R. J.; Balady, G. J.; Beasley, J. W., et al. ACC/AHA Guidelines for Exercise Testing. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Exercise Testing). Journal of the American College of Cardiology. 1997 Jul; 30(1):260-311"	NR-Review
2072	"Grilli, R.; Magrini, N.; Penna, A.; Mura, G., and Liberati, A. Practice guidelines developed by specialty societies: the need for a critical appraisal. Lancet. 2000 Jan 8; 355(9198):103-6"	NR-ROS
2074	"Guyatt, G. H.; DiCenso, A.; Farewell, V.; Willan, A., and Griffith, L. Randomized trials versus observational studies in adolescent pregnancy prevention. Journal of Clinical Epidemiology. 2000 Feb; 53(2):167-74"	NR-ROS
2078	"Hadorn, D. C.; Baker, D. W.; Kamberg, C. J., and Brooks, R. H. Phase II of the AHCPR-sponsored heart failure guideline: translating practice recommendations into review criteria. Joint Commission Journal on Quality Improvement. 1996 Apr; 22(4):265-76"	NR-Guideline
2080	"Haynes, R. B. Some problems in applying evidence in clinical practice. Annals of the New York Academy of Sciences. 1993 Dec 31; 703210-24; discussion 224-5"	NR-Review
2082	"Hedges, L. V. Improving meta-analysis for policy purposes. NIDA Research Monograph. 1997; 170:202-15"	NR-Stat Meth
2084	"Heisey, R.; Mahoney, L., and Watson, B. Management of palpable breast lumps. Consensus guideline for family physicians. Canadian Family Physician. 1999 Aug; 451926-32"	NR-Guideline
2088	"Hudak, P. L.; Cole, D. C., and Haines, A. T. Understanding prognosis to improve rehabilitation: the example of lateral elbow pain. Archives of Physical Medicine & Rehabilitation. 1996 Jun; 77(6):586-93"	NR-Not able to abstract
2092	"Irwig, L.; Zwarenstein, M.; Zwi, A., and Chalmers, I. A flow diagram to facilitate selection of interventions and research for health care. Bulletin of the World Health Organization. 1998; 76(1):17-24"	NR-OCD
2094	"Jacob, R. F., and Carr, A. B. Hierarchy of research design used to categorize the 'strength of evidence' in answering clinical dental questions. Journal of Prosthetic Dentistry. 2000 Feb; 83(2):137-52"	NR-Review
2096	"Jadad, A. R.; Cook, D. J., and Browman, G. P. A guide to interpreting discordant systematic reviews. Canadian Medical Association Journal. 1997 May 15; 156(10):1411-6"	NR-Review
2100	"Lipsey, M. W. Using linked meta-analysis to build policy models. NIDA Research Monograph. 1997; 170216-33"	NR-Modeling

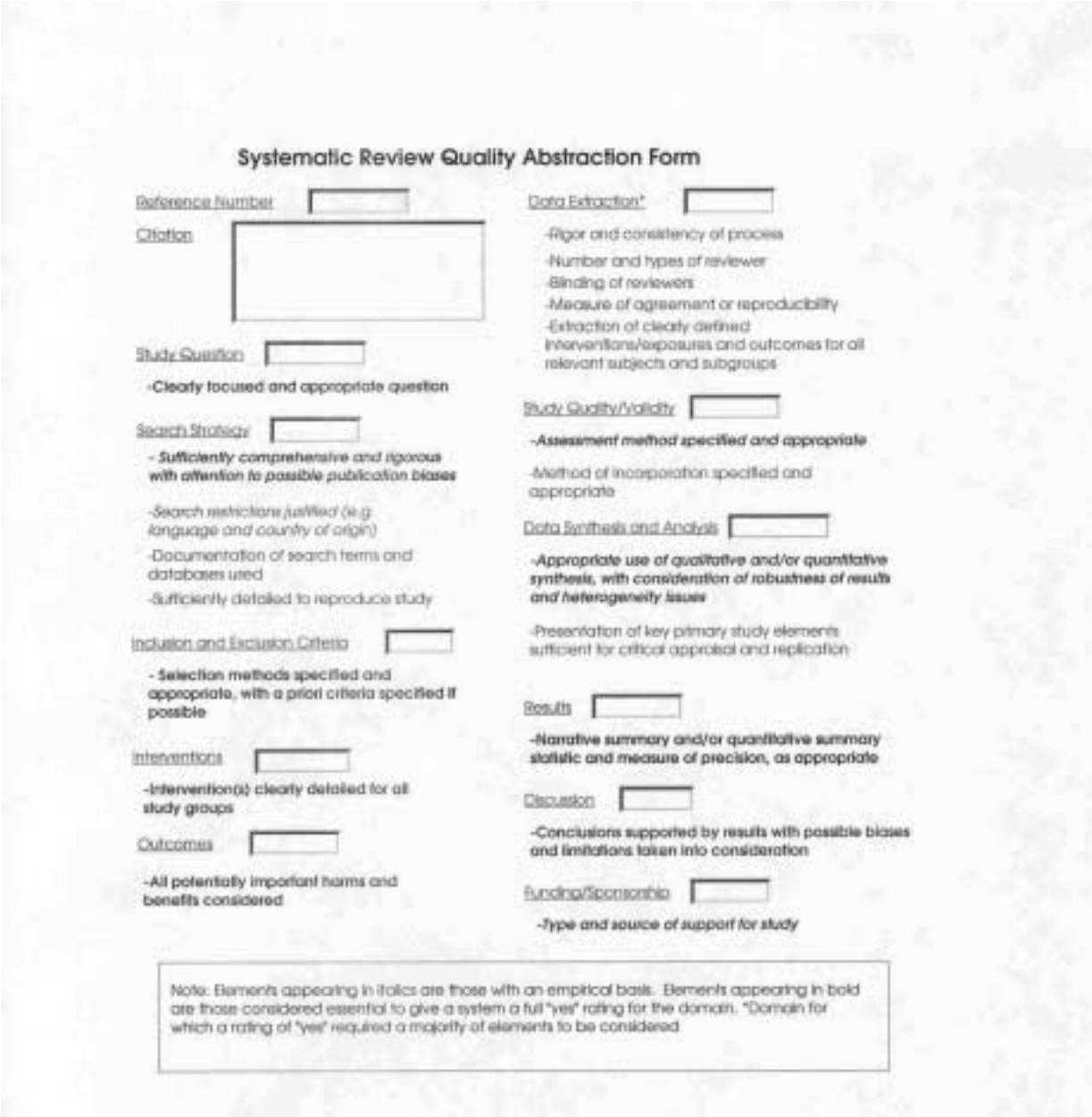
[*] See Table 4 for the code to abbreviations.

Excluded Articles (cont'd)

ID	Citation	Exclusion reason *[*]
2104	"Mackway-Jones, K.; Carley, S. D.; Morton, R. J., and Donnan, S. The best evidence topic report: a modified CAT for summarising the available evidence in emergency medicine. Journal of Accident & Emergency Medicine. 1998 Jul; 15(4):222-6"	NR-Review
2110	"Matt, G. E. Drawing generalized causal inferences based on meta-analysis. NIDA Research Monograph. 1997; 170165-82"	NR-Design/Methods
2116	"Newman, M. G., and McGuire, M. K. Evidence-based periodontal treatment. II. Predictable regeneration treatment. International Journal of Periodontics & Restorative Dentistry. 1995 Apr; 15(2):116-27"	NR-Design/Methods
2118	"Owens, D. K., and Nease, R. F. Jr. Development of outcome-based practice guidelines: a method for structuring problems and synthesizing evidence. Joint Commission Journal on Quality Improvement. 1993 Jul; 19(7):248-63"	NR-Modeling
2120	"Peterson, E. D.; Shaw, L. J., and Califf, R. M. Risk stratification after myocardial infarction [see comments]. [Review] [280 refs]. Annals of Internal Medicine. 1997 Apr 1; 126(7):561-82"	NR-Guideline
2124	"Ramirez, A. J.; Westcombe, A. M.; Burgess, C. C.; Sutton, S.; Littlejohns, P., and Richards, M. A. Factors predicting delayed presentation of symptomatic breast cancer: a systematic review. Lancet. 1999 Apr 3; 353(9159):1127-31"	NR-Not able to abstract
2126	"Schuster, M. A.; Asch, S. M.; McGlynn, E. A.; Kerr, E. A.; Hardy, A. M., and Gifford, D. S. Development of a quality of care measurement system for children and adolescents. Methodological considerations and comparisons with a system for adult women. Archives of Pediatrics & Adolescent Medicine. 1997 Nov; 151(11):1085-92"	NR-ROS
2128	"Stuck, A. E.; Walthert, J. M.; Nikolaus, T.; Bula, C. J.; Hohmann, C., and Beck, J. C. Risk factors for functional status decline in community-living elderly people: a systematic literature review. Social Science & Medicine. 1999 Feb; 48(4):445-69"	NR-ROS
2136	"Whitley, R. J.; Jacobson, M. A.; Friedberg, D. N.; Holland, G. N.; Jabs, D. A.; Dieterich, D. T.; Hardy, W. D.; Polis, M. A.; Deutsch, T. A.; Feinberg, J.; Spector, S. A.; Walmsley, S.; Drew, W. L.; Powderly, W. G.; Griffiths, P. D.; Benson, C. A., and Kessler, H. A. Guidelines for the treatment of cytomegalovirus diseases in patients with AIDS in the era of potent antiretroviral therapy: recommendations of an international panel. International AIDS Society-USA. Archives of Internal Medicine. 1998 May 11; 158(9):957-69"	NR-Guideline
2142	"Williams, D. N.; Rehm, S. J.; Tice, A. D.; Bradley, J. S.; Kind, A. C., and Craig, W. A. Practice guidelines for community-based parenteral anti-infective therapy. ISDA Practice Guidelines Committee. Clinical Infectious Diseases. 1997 Oct; 25(4):787-801"	NR-Guideline
2144	"Wilson, L. M.; Reid, A. J.; Midmer, D. K.; Biringer, A.; Carroll, J. C., and Stewart, D. E. Antenatal psychosocial risk factors associated with adverse postpartum family outcomes. Canadian Medical Association Journal. 1996 Mar 15; 154(6):785-99"	NR-Review

[*] See Table 4 for the code to abbreviations.

Appendix F: Abstraction Forms



Systematic Review Quality Abstraction Form

Reference Number

Citation

Study Question

-Clearly focused and appropriate question

Search Strategy

- Sufficiently comprehensive and rigorous with attention to possible publication biases

-Search restrictions justified (e.g. language and country of origin)

-Documentation of search terms and databases used

-Sufficiently detailed to reproduce study

Inclusion and Exclusion Criteria

- Selection methods specified and appropriate, with a priori criteria specified if possible

Interventions

-Intervention(s) clearly detailed for all study groups

Outcomes

-All potentially important harms and benefits considered

Data Extraction*

-Rigor and consistency of process

-Number and types of reviewer

-Blinding of reviewers

-Measure of agreement or reproducibility

-Extraction of clearly defined interventions/exposures and outcomes for all relevant subjects and subgroups

Study Quality/Validity

-Assessment method specified and appropriate

-Method of incorporation specified and appropriate

Data Synthesis and Analysis

-Appropriate use of qualitative and/or quantitative synthesis, with consideration of robustness of results and heterogeneity issues

-Presentation of key primary study elements sufficient for critical appraisal and replication

Results

-Narrative summary and/or quantitative summary statistic and measure of precision, as appropriate

Discussion

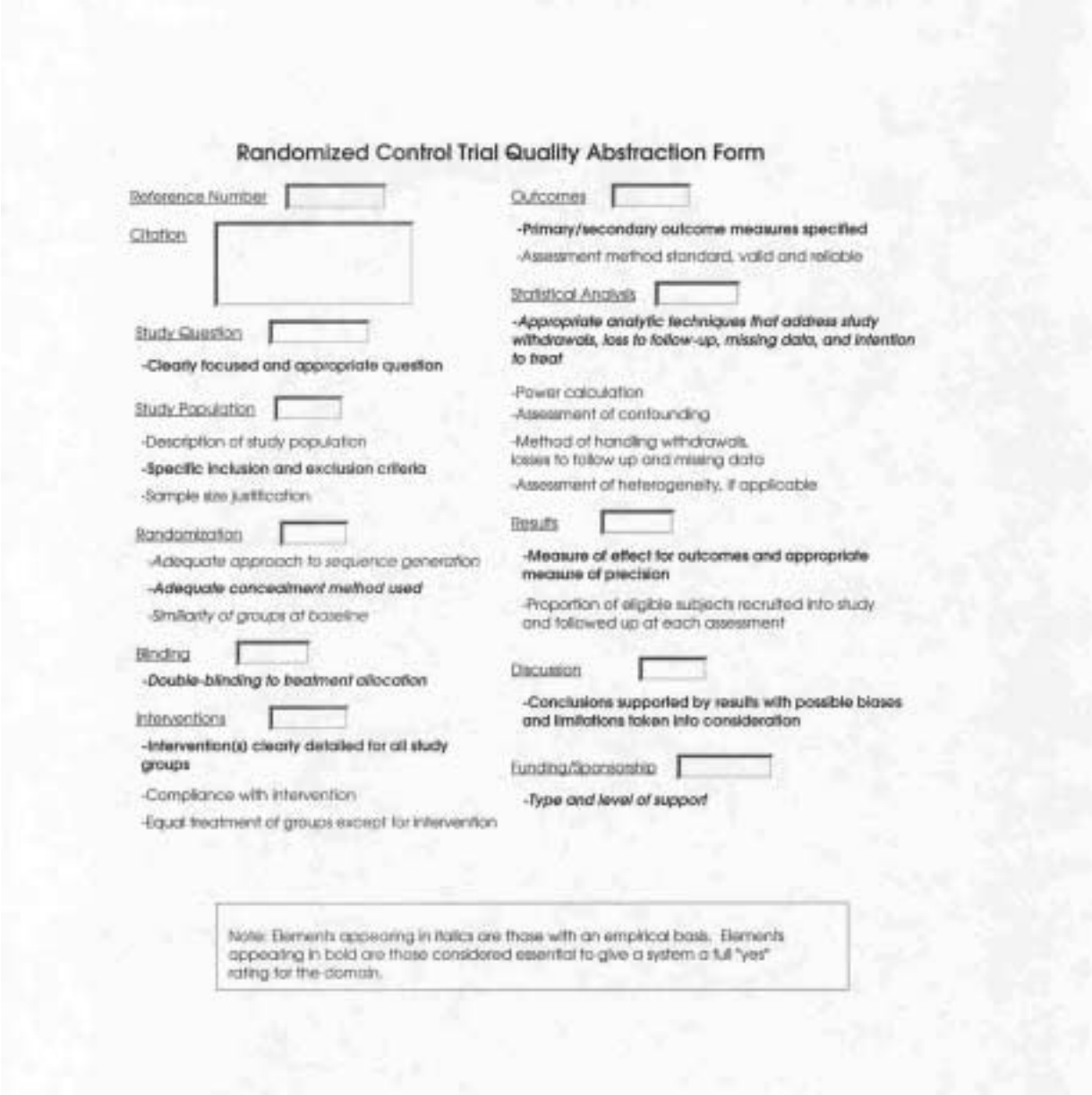
-Conclusions supported by results with possible biases and limitations taken into consideration

Funding/Sponsorship

-Type and source of support for study

Note: Elements appearing in italics are those with an empirical basis. Elements appearing in bold are those considered essential to give a system a full "yes" rating for the domain. *Domain for which a rating of "yes" required a majority of elements to be considered

Abstraction Forms (cont'd)



Randomized Control Trial Quality Abstraction Form

Reference Number

Citation

Study Question

-Clearly focused and appropriate question

Study Population

-Description of study population

-Specific inclusion and exclusion criteria

-Sample size justification

Randomization

-Adequate approach to sequence generation

-Adequate concealment method used

-Similarity of groups at baseline

Blinding

-Double-blinding to treatment allocation

Interventions

-Intervention(s) clearly detailed for all study groups

-Compliance with intervention

-Equal treatment of groups except for intervention

Outcomes

-Primary/secondary outcome measures specified

-Assessment method standard, valid and reliable

Statistical Analysis

-Appropriate analytic techniques that address study withdrawals, loss to follow-up, missing data, and intention to treat

-Power calculation

-Assessment of confounding

-Method of handling withdrawals, losses to follow up and missing data

-Assessment of heterogeneity, if applicable

Results

-Measure of effect for outcomes and appropriate measure of precision

-Proportion of eligible subjects recruited into study and followed up at each assessment

Discussion

-Conclusions supported by results with possible biases and limitations taken into consideration

Funding/Sponsorship

-Type and level of support

Note: Elements appearing in italics are those with an empirical basis. Elements appearing in bold are those considered essential to give a system a full "yes" rating for the domain.

Abstraction Forms (cont'd)

Observational Study Quality Abstraction Form

Reference Number

Citation

Study Question

-Clearly focused and appropriate question

Study Population

-Description of study populations

-Sample size justification

Comparability of subjects *

For all observational studies:

-Specific inclusion/exclusion criteria for all groups

-Criteria applied equally to all groups

-Comparability of groups at baseline with regard to disease status and prognostic factors

-Study groups comparable to non-participants with regard to confounding factors

-Use of concurrent controls

-Comparability of follow-up among groups at each assessment

Additional criteria for case-control studies:

-Explicit case definition

-Case ascertainment not influenced by exposure status

-Controls similar to cases except without condition of interest and with equal opportunity for exposure

Exposure/Intervention

-Clear definition of exposure

-Measurement method standard, valid and reliable

-Exposure measured equally in all study groups

Outcome Measurement

-Primary/secondary outcomes clearly defined

-Outcomes assessed blind to exposure or intervention status

-Method of outcome assessment standard, valid and reliable

-Length of follow-up adequate for question

Statistical analysis

-Statistical tests appropriate

-Multiple comparisons taken into consideration

-Modeling and multivariate techniques appropriate

-Power calculation provided

-Assessment of confounding

-Dose-response assessment, if appropriate

Results

-Measure of effect for outcomes and appropriate measure of precision

-Adequacy of follow-up for each study group

Discussion

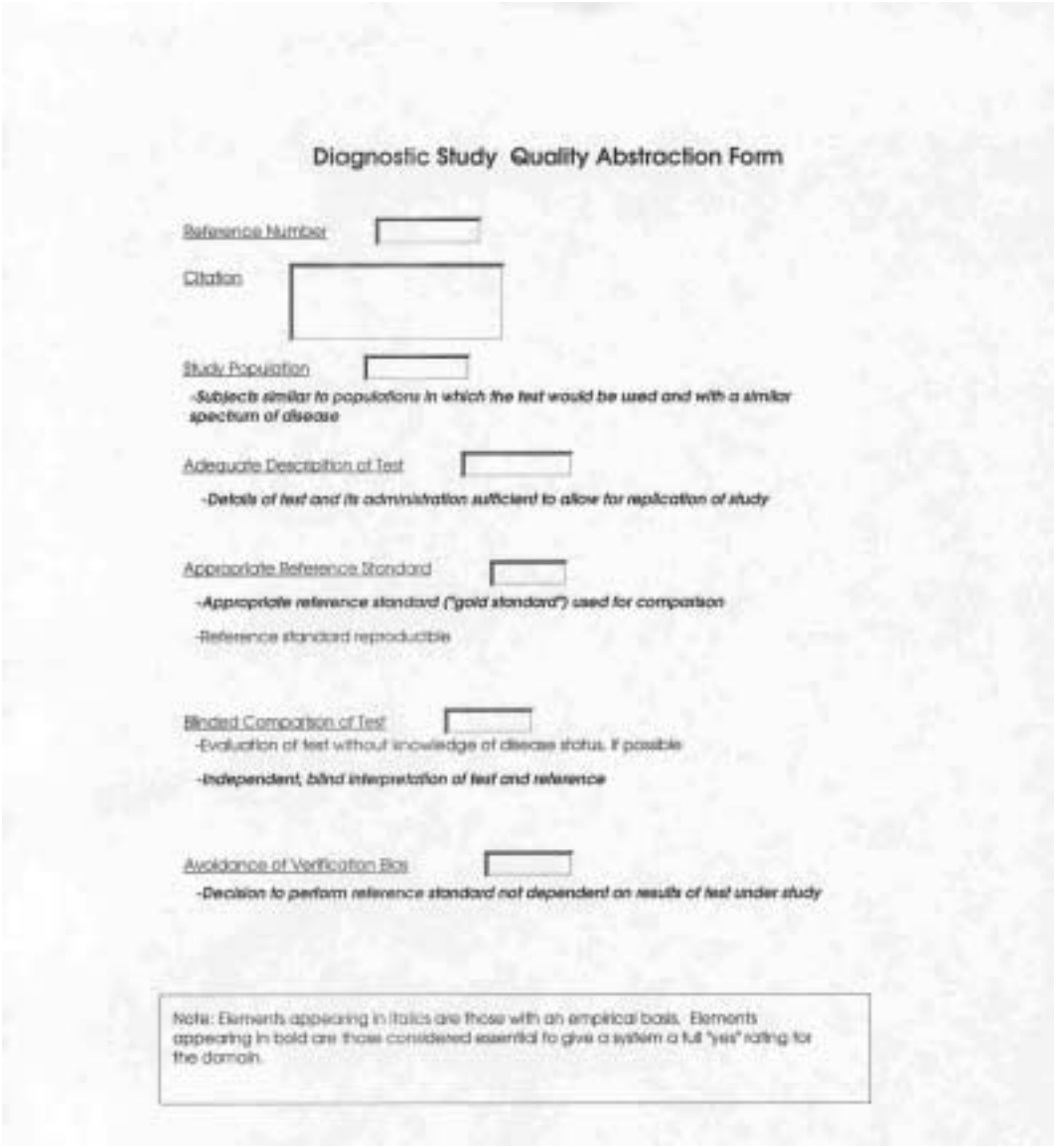
-Conclusions supported by results with biases and limitations taken into consideration

Funding/Sponsorship

-Type and sources of support for study

Note: Elements appearing in italics are those with an empirical basis. Elements appearing in bold are those considered essential to give a system a full "yes" rating for the domain. * Domain for which a rating of "yes" required a majority of elements be considered.

Abstraction Forms (cont'd)



Diagnostic Study Quality Abstraction Form

Reference Number

Citation

Study Population

-Subjects similar to populations in which the test would be used and with a similar spectrum of disease

Adequate Description of Test

-Details of test and its administration sufficient to allow for replication of study

Appropriate Reference Standard

-Appropriate reference standard ("gold standard") used for comparison

-Reference standard reproducible

Blinded Comparison of Test

-Evaluation of test without knowledge of disease status, if possible

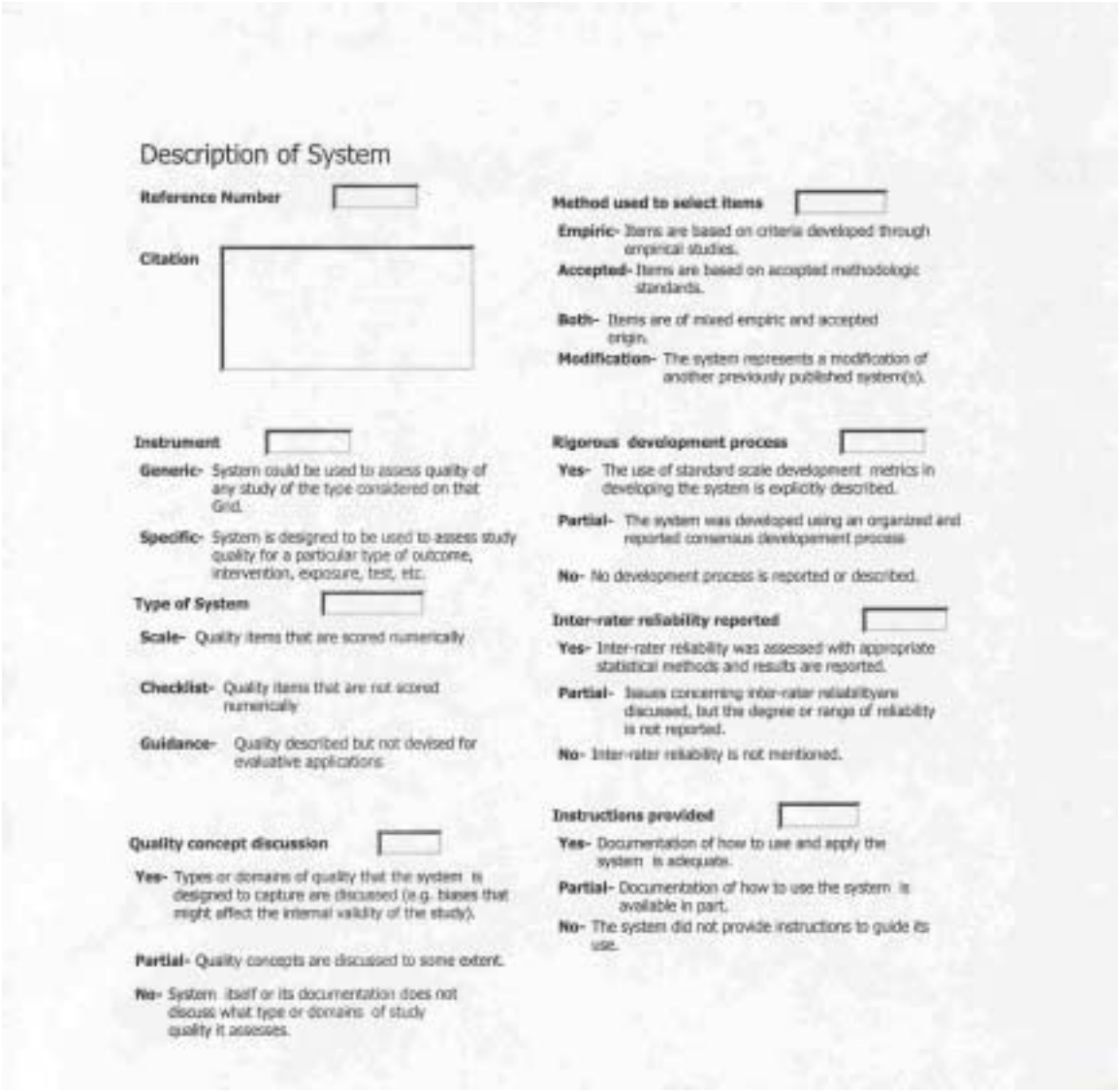
-Independent, blind interpretation of test and reference

Avoidance of Verification Bias

-Decision to perform reference standard not dependent on results of test under study

Note: Elements appearing in italics are those with an empirical basis. Elements appearing in bold are those considered essential to give a system a full "yes" rating for the domain.

Abstraction Forms (cont'd)



Description of System

Reference Number

Citation

Instrument

Generic- System could be used to assess quality of any study of the type considered on that Grid.

Specific- System is designed to be used to assess study quality for a particular type of outcome, intervention, exposure, test, etc.

Type of System

Scale- Quality items that are scored numerically

Checklist- Quality items that are not scored numerically

Guidance- Quality described but not devised for evaluative applications

Quality concept discussion

Yes- Types or domains of quality that the system is designed to capture are discussed (e.g. biases that might affect the internal validity of the study).

Partial- Quality concepts are discussed to some extent.

No- System itself or its documentation does not discuss what type or domains of study quality it assesses.

Method used to select items

Empiric- Items are based on criteria developed through empirical studies.

Accepted- Items are based on accepted methodologic standards.

Both- Items are of mixed empiric and accepted origin.

Modification- The system represents a modification of another previously published system(s).

Rigorous development process

Yes- The use of standard scale development metrics in developing the system is explicitly described.

Partial- The system was developed using an organized and reported consensus development process

No- No development process is reported or described.

Inter-rater reliability reported

Yes- Inter-rater reliability was assessed with appropriate statistical methods and results are reported.

Partial- Issues concerning inter-rater reliabilityare discussed, but the degree or range of reliability is not reported.

No- Inter-rater reliability is not mentioned.

Instructions provided

Yes- Documentation of how to use and apply the system is adequate.

Partial- Documentation of how to use the system is available in part.

No- The system did not provide instructions to guide its use.

Appendix G: Glossary

Abstraction	The method by which reviewers or researchers read scientific articles and then collect and record data from them.
AHRQ	Agency for Healthcare Research and Quality.
Allocation concealment	The processes used to prevent knowledge of group assignment in a randomized controlled trial before the actual intervention/treatment/exposure is administered. This process should be seen as distinct from blinding or masking of treatment group after the allocation process. The allocation process should be impervious to any influence by the individual making the allocation by having the randomization process administered by someone who is not responsible for recruiting participants.
Bias	Any systematic error in the design, conduct, or analysis of a study that results in a mistaken estimate of effect.
Case-control study	A type of observational study. Patients who have developed a disease or condition are identified and their past exposure to suspected etiological factors is compared with that of controls or referents who do not have the disease or condition.
The Cochrane Library©	An electronic publication of The Cochrane Collaboration, an international group dedicated to preparing, maintaining, and promoting the accessibility of systematic reviews of the effects of health care interventions.
Cohort study	A type of observational study. Factors related to the development of disease are measured initially in a group of persons, known as a cohort. The group is followed over a period of time and the relationship of a factor to the disease is examined. The population may be divided into subgroups according to the level or presence of the factor initially and comparing the subsequent incidence of disease in each subgroup.
Cohort	A subset of a population with a common feature, such as age, sex, or occupation.
Consistency	For any given topic, the extent to which similar findings are reported using similar or different study designs.

CONSORT	Consolidated Standards of Reporting Trials. A checklist of guidelines and items to be addressed when preparing published reports of RCTs.
Controls	A group of study subjects with whom a comparison is made in an epidemiologic study. For example, in a case-control study, cases are persons who have the disease and controls are persons who do not have the disease.
Diagnostic study	A study that examines the sensitivity and specificity of a particular test to evaluate to presence and/or absence of disease.
Domain	A quality construct relating to some aspect of study design or conduct considered important in determining the extent to which a study is valid.
Empirical	A concept designating that work is based directly on observational or experimental study, rather than theory or reasoning alone.
EPC	AHRQ Evidence-based Practice Center.
External validity	The extent to which a study can produce unbiased inferences regarding a target population (beyond the subjects of the study).
Gray literature	Materials that are found in recorded, written, or electronic form that are not traditionally well indexed or readily available. Examples are conference papers, white papers, technical reports, electronic theses and dissertations, online documents, and oral presentations/abstracts.
Guidance document	Publication that defines or describes study quality, but does not provide an instrument that could be used for evaluative applications.
Guidelines	Recommendations or principles presenting current or future guidance of policy, practice, or procedure. Guidelines are developed by government agencies at any level—institutions, professional societies, governing boards—or by the convening of expert panels. The formal definition of "clinical practice guidelines" comes from a 1990 report from the Institute of Medicine: "PRACTICE GUIDELINES are systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances."
Internal validity	The extent to which a study describes the "truth." A study conducted in a rigorous manner such that the observed differences between the experimental or observational groups and the outcomes under study may be attributed only to the hypothesized effect under investigation.
Inter-rater reliability	A measure of the extent to which multiple raters or judges agree when providing a rating, scoring, or assessment.

Magnitude of effect	The size or strength of the estimated association or effect observed in a given study. Magnitude of effect is often expressed as a odds ratio (OR) or relative risk (RR).
MEDLINE©	A comprehensive database, updated weekly, of bibliographic materials containing nearly 11 million records from more than 7,300 publications from 1965. It is compiled by the U.S. National Library of Medicine (NLM) and published on the Web by Community of Science.
Meta-analysis	The process of using statistical methods to combine quantitatively the results of similar studies in a systematic review.
Methodology	The scientific study of methods, or the practices and procedures used to plan, conduct, and analyze the results of a scientific study.
MOOSE	Meta-analysis Of Observational Studies in Epidemiology. A consensus workshop held in Atlanta, Georgia, in April 1997, convened by the Centers for Disease Control and Prevention, to examine the reporting of meta-analyses of observational studies and to make recommendations.
Peer-reviewed literature	Publications including research proposals, manuscripts submitted for publication, and abstracts submitted for presentation at scientific meetings that are judged for scientific and technical merit by other scientists in the same field.
Prospective cumulative meta-analysis	A meta-analysis that is conducted by adding each new study's results on a particular topic as it is available.
Quality checklists	Instruments that contain a number of quality items, none of which is scored numerically.
Quality component	Individual aspect of study methodology—for example, randomization, blinding, follow-up—that has a potential relation to bias in estimation of effect.
Quality scales	Instruments that contain several quality items that are scored numerically to provide a quantitative estimate of overall study quality.
QUORUM	The Quality of Reporting of Meta-Analyses. A QUORUM statement, checklist, and flow diagram stemming from a conference to address

	standards for improving the quality of reporting of meta-analyses of randomized controlled trials.
Randomization	The process of allocating a particular experimental intervention or exposure to a group at random, in order to control for all other factors that may affect disease risk.
Randomized clinical trial (RCT)	A clinical trial that involves at least one treatment and one control group, concurrent enrollment, and follow-up of the groups, and in which the treatments to be allocated are selected by a random process, such as the use of a random numbers table.
Retrospective cohort study	A type of observational study. This study design begins with a group of affected individuals and tests the hypothesis that some prior characteristic or exposure is more common in persons with the disease than in unaffected persons.
Selection bias	Error attributable to systematic differences in characteristics between those who are selected for study and those who are not.
Sensitivity	The proportion of truly diseased persons in the screened population who are identified as diseased by the screening test—that is, the true-positive rate.
Sensitivity analysis	Determining the robustness of analysis by examining the extent to which changes in methods, values of variables, or assumptions change results. The aim is to identify variables whose values are most likely to alter results or to find a solution that is relatively stable for the commonly occurring values of these variables.
Specificity	The proportion of truly nondiseased persons who are identified as such by the screening test—that is, the true-negative rate.
STARD	STAndards for Reporting Diagnostic Accuracy. Developed by an international group addressing the need for quality measures for studies of diagnostic services.
Statistical power	The statistical ability of a study to correctly identify a true difference between therapies. Power chiefly depends upon the number of subjects in a study and the response rate of the study groups.
Systematic review	An organized method of locating, assembling, and evaluating a body of literature on a particular topic using a set of specific predefined criteria. A

	systematic review may be purely narrative or may also include a quantitative pooling of data, referred to as a meta-analysis.
TEAG	Technical Expert Advisory Group
Temporality	The relationship of time and events such as exposure to a risk factor and the development of disease. To implicate the exposure as causative of the disease, the exposure should have occurred before the disease.