Examining the Evidence

A series presenting finding from a systematic search of the literature on a specific topic and offering quantitative or qualitative analysis of these findings.

Rating the strength of scientific evidence: relevance for quality improvement programs

KATHLEEN N. LOHR
RTI International, Research Triangle Park, NC, USA

Abstract

Objectives. To summarize an extensive review of systems for grading the quality of research articles and rating the strength of bodies of evidence, and to highlight for health professionals and decision-makers concerned with quality measurement and improvement the available ‘best practices’ tools by which these steps can be accomplished.

Design. Drawing on an extensive review of checklists, questionnaires, and other tools in the field of evidence-based practice, this paper discusses clinical, management, and policy rationales for rating strength of evidence in a quality improvement context, and documents best practices methods for these tasks.

Results. After review of 121 systems for grading the quality of articles, 19 systems, mostly study design specific, met a priori scientific standards for grading systematic reviews, randomized controlled trials, observational studies, and diagnostic tests; eight systems (of 40 reviewed) met similar standards for rating the overall strength of evidence. All can be used as is or adapted for particular types of evidence reports or systematic reviews.

Conclusions. Formally grading study quality and rating overall strength of evidence, using sound instruments and procedures, can produce reasonable levels of confidence about the science base for parts of quality improvement programs. With such information, health care professionals and administrators concerned with quality improvement can understand better the level of science (versus only clinical consensus or opinion) that supports practice guidelines, review criteria, and assessments that feed into quality assurance and improvement programs. New systems are appearing and research is needed to confirm the conceptual and practical underpinnings of these grading and rating systems, but the need for those developing systematic reviews, practice guidelines, and quality or audit criteria to understand and undertake these steps is becoming increasingly clear.

Keywords: clinical practice guidelines, evidence-based practice, quality improvement, quality of care strength of evidence

Introduction

Around the globe, a ‘trend to evidence’ appears to motivate the search for answers to markedly disparate questions about the costs and quality of health care, access to care, risk factors for disease, social determinants of health, and indeed about the air we breathe and the food we eat. We look for solutions to problems of rare or genetic disorders, seek guidance on the safest, most effective treatments for everything from the common cold to childhood cancers, and expect to be informed about the ‘best’ (or ‘worst’) hospitals and doctors in our cities and towns. The call is strong for science to help stave off premature death, needless disability, and wasteful expenditures of personal or government money.

In making informed choices about health care, people increasingly seek credible evidence. Such evidence reflects ‘empirical observations…of real events, [that is,] systematic observations using rigorous experimental designs or non-systematic observations (e.g. experience)…not revelations, dreams, or ancient texts'[1]. For situations as different as clinical care, policy-making, dispute resolution, and law [2,3], evidence needs to be seen as both relevant and reliable; science and collected
bodies of evidence, however, need to be tempered by clinical acumen and political realities. In addressing issues of the quality of health care ‘the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge’ ([4], p. 21) this mix of science and art is crucial.

Quality assessment and improvement activities rest heavily on clinical practice guidelines (CPGs) and review and audit criteria. CPGs (‘systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances’ [5], p. 27) can improve health professionals’ knowledge by providing information and recommendations about appropriate and needed services for all aspects of patient management: screening and prevention, diagnosis, treatment, rehabilitation, palliation, and end-of-life care. When kept updated as technologies change, CPGs also influence attitudes about standards of care and, over time, shift practice patterns to make care more efficient and effective, thereby enhancing the value received for health care outlays. Moreover, evidence-based guidelines constitute a major element of quality assurance, quality improvement, medical audit, and similar activities for many health care settings: inpatient or residential (e.g. hospitals, nursing homes), outpatient (e.g. offices, ambulatory clinics, and private homes), and emergency departments or clinics. Users can convert them into medical review criteria to assess care generally and in these settings or to target specific kinds of services, providers, settings, or patient populations for in-depth review [2,6].

Evidence-based practice brings pertinent, trustworthy information into this equation by systematically acquiring, analyzing, and transferring research findings into clinical, management, and policy arenas. The process involves:

• developing the question in a way that can be answered by a systematic review: specifying the populations, settings, problems, interventions, and outcomes of interest;
• stating criteria for eligibility (inclusion and exclusion) of literature to be considered before conducting literature searches, so as to avoid bias introduced by arbitrarily including or excluding certain studies;
• searching the literature to capture all the evidence about the question of interest;
• reviewing abstracts of publications to determine initial eligibility of studies;
• reviewing retained studies to determine final eligibility;
• abstracting data on these studies into evidence tables;
• determining the quality of studies and the overall strength of evidence;
• synthesizing and combining data from evidence tables, and deciding whether quantitative analyses (i.e. meta-analysis) are warranted; and
• writing a draft review, subjecting it to peer review, editing and revising, and producing the final review.

This paper examines one evidence-based process—rating the quality and strength of evidence—to argue three points:

(1) The confidence that those wishing to mount credible quality improvement (QI) efforts can assign to evidence rests in part on the quality of individual research efforts and the overall strength of those bodies of evidence; with such assurance, they can distinguish more clearly between good and bad information and between evidence and mere opinion.

(2) Formal efforts to grade study quality and rate the strength of evidence can produce a reasonable level of confidence about that evidence.

(3) Tools that meet acceptable scientific standards can facilitate these grading and rating steps.

Evidence and evidence-based practice

Evidence-based practice

Evidence-based medicine is ‘the integration of best research evidence with clinical expertise and patient values’ [7]. In clinical applications, providers use the best evidence available to decide, together with their patients, on suitable options for care. Such evidence comes from different types of studies conducted in various patient groups or populations. The emphasis is on melding scientific evidence of the highest caliber with sensitive appreciation of patients’ values and preferences—blending the science and art of medicine.

One challenge for practitioners is that most medical recommendations today refer to groups of patients (‘women over age 50’), and they may or may not apply to a particular woman with a particular medical history and set of cultural values. Moreover, when evidence for an intervention is relatively weak, e.g. benefits and harms of prostate-specific antigen screening for prostate cancer [8] or the value of universal screening of newborns for hearing loss to improve long-term language outcomes [9], patients and providers are likely to give more emphasis to patients’ values and treatment costs. When evidence is strong, e.g. use of aspirin to prevent heart attacks, especially in high-risk patients [10], the value of screening for colorectal cancer [11], or the payoff from stopping smoking [12], patients’ values may carry less weight in treatment decisions, although their preferences for different outcomes always need to be taken into account.

Even though health care management and administration is moving into an evidence-based environment (see for example Evidence-Based Healthcare, available at http://www.hbuk.co.uk/journals/ebhe), executives concerned with implementing proven or innovative QI programs face similar challenges. Numerous for-profit and non-profit organizations help hospitals, group practices, delivery systems, and large health plans implement and evaluate approaches to change organizational structures and behaviors to improve clinical and patient outcomes, enhance patient safety, attain better cost and cost-effectiveness goals, and address the ‘business case for quality’ question [13]. Other enterprises create evidence-based prescription information tools and web content with consumer health information. Yet other institutions focus on practice guidelines (e.g. http://www.guidelines.gov; http://medicine.ucsf.edu/resources/guidelines). In Europe, BIOMED-supported activities are a related effort to develop a tool for assessing
Rating scientific evidence

Inventories of process and outcome measures add yet another dimension to these activities (http://www.qualitymeasures.ahrq.gov). Faster adoption of useful innovations, including QI programs, is seen as a particularly critical endeavor [14]. In all these arenas, sound evidence is critical.

Evidence-based recommendations that take into account benefits and harms of health interventions give those responsible for QI planning and decisions grounds for adopting some technologies or programs and abandoning others, although the proposition that research can have a direct influence on such decision-making can be questioned [15–18]. The next frontier may lie in finding ways to organize knowledge bases better, or to set up independent centers or other efforts to support data collection, research, analysis, and modeling specifically pertinent to QI programs [19–22].

The nature of desirable evidence

QI programs need information across the entire spectrum of biomedical, clinical, and health services research. Good evidence, applicable to all patients and care settings, is not available for much of medicine today. Perhaps no more than half, or even one-third, of services are supported by compelling evidence that benefits outweigh harms. Millenson claims, citing work from Williamson in the late 1970s [23], that ‘[m]ore than half of all medical treatments, and perhaps as many as 85 percent, have never been validated by clinical trials’ ([24], p. 15). According to an expert committee of the US Institute of Medicine, only about 4% of all services have strong strength of evidence and modest to strong clinical consensus and more than 50% of services had very weak or no evidence ([5], Tables 1 and 2). Although clinical and health services research have escalated in the intervening years, so has the technological armamentarium and spectrum of disease, suggesting major gaps still remain for research to fill and that major challenges lie ahead for the development of systematic reviews on clinical and health care delivery topics.

In this context, the absence of evidence about benefits (or harms) is not the same as evidence of no benefit (or harm). For deciding whether to render a medical service or cover a new technology, clinicians, administrators, guideline developers, and even patients must be alert to this distinction. ‘No evidence’ is a reason for caution in reaching judgments and clinical or policy decisions and for postponing definitive steps. In contrast, ‘evidence of no positive (or negative) impact’ may be a solid reason for taking conclusive steps in favor of or against a medical service.

Evidence, even when available, is rarely definitive. The level of confidence that one might have in evidence turns on

<table>
<thead>
<tr>
<th>Systematic reviews</th>
<th>Randomized controlled trials</th>
<th>Observational studies</th>
<th>Diagnostic test studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study question</td>
<td>Study question</td>
<td>Study question</td>
<td>Study question</td>
</tr>
<tr>
<td>Search strategy</td>
<td>Study population</td>
<td>Study population</td>
<td>Study population</td>
</tr>
<tr>
<td>Inclusion and exclusion criteria</td>
<td>Randomization</td>
<td>Comparability of subjects</td>
<td>Adequate description of test</td>
</tr>
<tr>
<td>Interventions</td>
<td>Blinding</td>
<td>Exposure or intervention</td>
<td>Appropriate reference standard</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Interventions</td>
<td>Outcome measures</td>
<td>Blinded comparison of test and standard</td>
</tr>
<tr>
<td>Data extraction</td>
<td>Statistical analysis</td>
<td>Statistical analysis</td>
<td>Avoidance of verification bias</td>
</tr>
<tr>
<td>Study quality and validity</td>
<td>Results</td>
<td>Results</td>
<td>Discussion</td>
</tr>
<tr>
<td>Data synthesis and analysis</td>
<td>Discussion</td>
<td>Discussion</td>
<td>Funding or sponsorship</td>
</tr>
<tr>
<td>Results</td>
<td>Funding or sponsorship</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Funding or sponsorship</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: West et al. (2002) [26].

Italics indicate elements of critical importance in evaluating grading systems according to empirical validation research or standard epidemiological methods.

| Table 2 Criteria for evaluating systems to rate the strength of bodies of evidence |
|----------------------------------------|---------------------------------|-----------------------------|------------------------|
| Quality | The aggregate of quality ratings for individual studies, predicated on the extent to which bias was minimized |
| Quantity | Numbers of studies, sample size or power, and magnitude of effect |
| Consistency | For any given topic, the extent to which similar findings are reported using similar and different study designs |

Source: West et al. (2002) [26].
the underlying robustness of the research and the analyses done to synthesize that research. Users can, and of course often do, arrive at their own judgments about the soundness of practice guidelines or technology assessments and the science underpinning their conclusions and recommendations. Such judgments may differ considerably in the sophistication and lack of bias with which they were made, for any number of reasons: disputing which evidence is appropriate for assessment in the first place; examining only some of the evidence; disagreeing as to whether factors such as patient satisfaction and cost should be explicitly included in the assessment of the effectiveness of a diagnostic test or treatment; and differing in conclusions about the quality of the evidence. Without consensus on what constitutes sufficient evidence of acceptable quality, such disagreement is not surprising, but it can lead to public concern either that the evidence on many issues is ‘bad’ or that the experts somehow represent a collection of special interests and ought not wholly to be trusted.

For that reason, groups producing systematic reviews, as the underpinnings to guidelines or quality and audit review criteria, are likely to be in the best position to evaluate the strength of the evidence they are assembling and analyzing. Nonetheless, they must be transparent about how they reached such judgments in the first place. Explicitly evaluating the quality of research studies and judging the strength of bodies of evidence is a central, inseparable part of this process.

Grading quality and rating the strength of evidence

Defining quality and strength in evidence-based practice terms

Grading the quality of individual studies and rating the strength of the body of evidence comprising those studies are the two linked topics for the remainder of this paper. Quality, in this context, is ‘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’ ([25], p. 472). An expanded view holds that quality concerns the extent to which a study’s design, conduct, and analysis have minimized biases in selecting subjects and measuring both outcomes and differences in the study groups other than the factors being studied that might influence the results [26].

In practical terms, one can grade studies only by examining the details that articles in the peer-reviewed literature provide. If studies are incompletely or inaccurately documented, they are likely to be downgraded in quality (perhaps fairly, perhaps not). New guidelines from international groups provide clear instructions on how systematic reviews (QUORUM), randomized controlled trials (CONSORT), observational studies (MOOSE), and studies of diagnostic test accuracy (STARD) ought to be reported [27–30]. These statements are not, however, direct tools for evaluating the quality of studies.

Strength of evidence has a similar range of definitions, all taking into account the size, credibility, and robustness of the combined studies on a given topic. It ‘incorporates judgments of study quality [and] 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’ [26]. ‘Closeness to the truth’, ‘size of the effect’, and ‘applicability (usefulness in . . . clinical practice)’ are the concepts used by some evidence-based experts to convey the idea of strength of evidence [7].

The US Preventive Services Task Force, for example, holds that the strength of evidence applies to linkages in an analytic framework for a clinical question that might run from screening to confirmatory diagnosis, treatment, intermediate outcomes (e.g. biophysical measures), and ultimately patient outcomes (e.g. survival, functioning, emotional well-being, and satisfaction) [31]. Criteria for judging evidentiary strength involve internal validity (the extent to which studies yield valid information about the populations and settings in which they were conducted), external validity (the extent to which studies are relevant and can be generalized to broader patient populations of interest), and coherence or consistency (the extent to which the body of evidence makes sense, given the underlying model for the clinical situation).

Strength of evidence needs to be distinguished from the magnitude of effect or impact reported in research papers. How solid we believe a body of evidence is ought not to be confused with how dramatic the effects and outcomes have been. Very robust evidence in favor of small effects of clinical interventions may prove more telling in QI decision-making than weak evidence about ostensibly spectacular findings. Cutting across these considerations is the frequency or rarity of benefits or harms. Holding the amount or explanatory power of the evidence constant, weighing common small benefits against rare but catastrophic harms is a difficult, and sometimes subjective, tradeoff.

Both conceptually and practically, quality and strength are related, albeit hierarchical, ideas. One must grade the quality of individual studies before one can draw affirmative conclusions about the strength of the aggregated evidence. These steps feed directly into grading health care recommendations relevant to QI programs.

Although this paper confines itself to study quality and strength of evidence, this link to assigning levels of confidence in recommendations is a straightforward and important one. For example, the USPSTF clearly explains its methods in a linked model that runs from grading studies to assessing strength of evidence to grading its recommendations [31]. GRADE is a new international effort related to reporting requirements that aims to develop a comprehensive approach to grading evidence and guideline recommendations (Andy Oxman, Norwegian Directorate for Health and Social Welfare, Oslo, personal communication, 6 May 2003).

In summary, grading studies and rating the strength of evidence matter because they can:

- clarify how certain one can be about research results and, thus, about conclusions, decisions, or recommendations drawn from that research;
- identify and perhaps alleviate problems of potential bias in the literature; and
• make transparent how taking quality of studies and strength of evidence into account affects aggregate findings and decisions to be made from those findings.

Methods

General approach

The US Agency for Healthcare Research and Quality (AHRQ) plays a significant role in evidence-based practice through its Evidence-based Practice Center (EPC) program and in quality of care [32]. In 1999, the US Congress directed AHRQ to examine systems to rate the strength of the scientific evidence underlying health care practices, research recommendations, and technology assessments and to make such methods or systems widely available. To fulfill this congressional charge, AHRQ commissioned the RTI International-University of North Carolina (RTI-UNC) EPC to produce an extensive evidence report that would: (i) describe systems that rate the quality of evidence in individual studies or grade the strength of entire bodies of evidence concerned with a single scientific question; and (ii) provide guidance on ‘best practices’ in this field today.

To complete this work required establishing criteria for judging systems for grading quality and rating strength of evidence, identifying such systems from the world literature and internet sites, evaluating the systems against these criteria, and judging which systems passed sufficient muster that they might be characterized as best practices. We conducted extensive literature searches of MEDLINE for articles published between 1995 and 2000 and sought further information from existing bibliographies, other sources including websites of several international organizations, and our expert panel advisers. In all, we reviewed 1602 publication abstracts. We developed and refined sets of evaluation criteria, which covered attributes and domains that reflect accepted principles of health research and epidemiology, relying on empirical research in the peer-reviewed literature and standard epidemiological texts. In addition, we relied extensively on members of an international technical panel comprising seasoned researchers and noted experts in evidence-based practice to provide feedback on our overall approach, including specification of our evaluation criteria. We developed and completed descriptive tables, similar to evidence tables, by which to compare and characterize existing systems, using the attributes and domains that we believed any acceptable instrument for these purposes ought to cover. After determining which grading and rating systems adequately covered the domains of interest (i.e. tools that fully or partially met the evaluation criteria), we identified those systems that we believed could be used more or less ‘as is’ (or easily adapted) and displayed this information in tabular form. These methods are described in detail elsewhere [26].

Grading study quality

For evaluating systems related to grading the quality of individual studies, the RTI-UNC EPC team defined domains for four types of research: systematic reviews (including ones that statistically combine data from individual studies), randomized controlled trials (RCTs), observational studies (which include a wide array of nonexperimental or quasi-experimental designs both with and without control or comparison groups), and investigations of diagnostic tests. As listed in Table 1, we specified both desirable domains and, of those, domains considered absolutely critical for a grading scheme to be regarded as acceptable (the latter are identified by italics). For example, for RCTs, adequate statement of the study question is a desirable domain that a grading scheme should cover, but adequate description of study population, randomization, and blinding are critical domains.

Rating strength of evidence

To evaluate schemes to rate the strength of a body of evidence, we specified three sets of aggregate criteria (Table 2) that combine key aspects of the design, conduct, and analysis of multiple studies on a given topic. The quality of evidence is essentially a summation of the direct grading of individual articles. The quantity of evidence concerns several variables that reflect the magnitude of effects (benefits and harms) estimated in these studies. Finally, the coherence or consistency of results reflects the extent to which studies report findings that reflect effects of similar magnitude and direction or that report discrepant findings that nonetheless can be explained adequately by biological, population, setting, or other characteristics.

Report preparation

The EPC team completed its evaluation and prepared a draft evidence report that was subjected to extensive external peer review, revised the report accordingly, and submitted the final to AHRQ. Subsequently, AHRQ organized a 1-day invitational conference of quality of care and other experts to discuss the ramifications of the report and avenues for dissemination to numerous audiences concerned with various aspects of health care delivery, including quality improvement. This paper was developed in response to the group’s general recommendations.

Results

Grading study quality

The EPC investigators assessed 121 grading systems against the domain-specific criteria specified a priori for systematic reviews, RCTs, observational studies, and diagnostic test studies and assigned scores of fully met, partially met, or not met (or no information). From these objective comparisons, the team classified 19 generic scales or checklists as ones that can be used in producing systematic evidence reviews, technology assessments, or other QI-related materials [33–51]. Tables 3a–3d depict the extent to which they met evaluation criteria.
Table 3a  Evaluation of systems to grade the quality of systematic reviews

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Critical domains in the evaluation criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Study question</td>
</tr>
<tr>
<td>Irwig et al. (1994) [51]</td>
<td>•</td>
</tr>
<tr>
<td>Sacks et al. (1996) [33]</td>
<td>•</td>
</tr>
<tr>
<td>Auperin et al. (1997) [34]</td>
<td>•</td>
</tr>
<tr>
<td>Barnes and Bero (1998) [35]</td>
<td>•</td>
</tr>
<tr>
<td>Khan et al. (2000) [36]</td>
<td>•</td>
</tr>
</tbody>
</table>

Legend: * = yes;  = partial; ° = not met or no information.
Source: West et al. (2002) [26].

Table 3b  Evaluation of systems to grade the quality of randomized controlled trials

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Critical domains in the evaluation criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Study population</td>
</tr>
<tr>
<td>Chalmers et al. (1981) [37]</td>
<td>•</td>
</tr>
<tr>
<td>Liberati et al. (1986) [38]</td>
<td>•</td>
</tr>
<tr>
<td>Reisch et al. (1989) [39]</td>
<td>•</td>
</tr>
<tr>
<td>van der Heijden et al. (1996) [40]</td>
<td>•</td>
</tr>
<tr>
<td>de Vet et al. (1997) [41]</td>
<td>•</td>
</tr>
<tr>
<td>Sindhu et al. (1997) [42]</td>
<td>•</td>
</tr>
<tr>
<td>Downs and Black (1998) [43]</td>
<td>•</td>
</tr>
<tr>
<td>Harbour and Miller (2001) [44]</td>
<td>•</td>
</tr>
</tbody>
</table>

1Instruments for RCTs only.
2Instruments for both RCTs and observational studies.
Source: West et al. (2002) [26].

Table 3c  Evaluation of systems to grade the quality of observational studies

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Critical domains in the evaluation criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comparability of subjects</td>
</tr>
<tr>
<td>Reisch et al. (1989) [39]</td>
<td>•</td>
</tr>
<tr>
<td>Spitzer et al. (1990) [45]</td>
<td>•</td>
</tr>
<tr>
<td>Goodman et al. (1994) [46]</td>
<td>•</td>
</tr>
<tr>
<td>Downs and Black (1998) [43]</td>
<td>•</td>
</tr>
<tr>
<td>Zaza et al. (2000) [47]</td>
<td>•</td>
</tr>
<tr>
<td>Harbour and Miller (2001) [44]</td>
<td>•</td>
</tr>
</tbody>
</table>

1Instruments for both RCTs and observational studies.
2Instruments for observational studies only.
Source: West et al. (2002) [26].
Rating strength of evidence

After evaluating 40 systems for rating strength against the quality, quantity, and consistency criteria, we identified eight instruments that fully addressed all three domains for rating the strength of a body of evidence (Table 4) [31,52–58]. The team also identified an additional nine approaches that incorporated three domains either fully or partially [7,36,44,59–64].

Discussion

Tools to draw on

Grading studies and rating strength of evidence can be done, and done well, with existing systems. For incorporating study quality and strength of evidence evaluations in systematic reviews, evidence reports, or technology assessments, groups can comfortably use one or more of these systems as a starting point. The EPC’s technical report describes and discusses the systems in more detail, because potential users need to take feasibility, ease of application, and certain other properties of these tools into account in selecting among them. The core conclusion remains: these systems constitute an acceptable set of tools available today for this critical step in developing products applicable to QI initiatives.

Agreement in principle about these ideas across scientists in several countries attests to the sturdiness of the core elements and concepts for assessing quality of studies and strength of evidence. Outcome measures, for example, are thought to be adequate when they are reliable (giving roughly the same answers when administered twice in short order), valid (measuring what they purport to measure), and clinically sensible. The factor of funding and sponsorship has been empirically validated more than once.

No one best approach

The EPC team offered other conclusions and observations about the state of the art, and science, of these tasks. Possibly most important is that there is no ‘best approach’. Acceptable methods for grading the quality of studies must take the original study design into account; approaches suitable for RCTs...
or observational studies will not be applicable for diagnostic tests, for instance. Even systems that are said to be applicable to both RCTs or observational research may prove to be difficult to use and yield less precise or reliable judgments than desired.

RCTs minimize selection bias, an important potential problem in observational studies. However, effectiveness and observational studies usually have larger total numbers of subjects and reflect more culturally, ethnically, and socially diverse patient populations and practice settings. No system for evaluating either quality or strength, no matter how good it seems to be, can completely resolve the inherent tension between these strengths (or weaknesses) of efficacy and effectiveness research. Users should match the topic and types of studies under review to an appropriate grading tool; one size will not fit all.

Future research, development, and evaluation

Even with these various rating and grading systems on the shelf, those in the QI world need to appreciate the work still needed to develop additional tools, provide better advice on how to use existing tools, and generate empirical documentation of the reliability and validity of new or extant systems. The extent to which these grading and rating steps influence guideline conclusions and recommendations needs to be evaluated. Until these research gaps are bridged, those wishing to produce authoritative systematic reviews, technology assessments, or QI and audit criteria will be hindered in their efforts. Future studies should: (i) address technical measurement issues; (ii) clarify the applicability of different systems to new, different, or less traditional clinical or policy topics; (iii) determine what factors make a difference in final quality scores for individual studies and, by extension, in judgments about the strength of bodies of evidence; and (iv) possibly most important, ascertain the impact of this process on conclusions, recommendations for QI programs, and ultimate health and policy outcomes.

Clinicians, managers, and QI leaders all face escalating demands on their time in an environment of increasingly complex decision-making like that reflected in Figure 1. Sorting out the science that enables practitioners, QI experts, and the public to make informed decisions is time-consuming and challenging substantively, given the accelerating pace of scientific discovery and production of peer- and non-peer-reviewed literature. They can turn to evidence-based systematic reviews, guidelines, and recommendations for help, but they must have confidence in this information base if they are to proceed with conviction and authority and if they are to be held accountable for the resulting clinical or policy choices they make.

Two critical tasks in developing defensible evidence-based reviews, which form the basis of practice guidelines, quality review and audit criteria, and similar materials, are to grade the quality of individual studies and then to rate the strength of the overall body of evidence. When evidence-based reviews and recommendations incorporate these steps, decision-makers from the national policy level to the individual physician–patient relationship can have greater assurance that their choices will be well-informed, well-grounded, and appropriate to the challenges ahead.

References


Accepted for publication 24 September 2003