COMMENTARY

AHRQ Series Commentary 1: Rating the evidence in comparative effectiveness reviews

Yngve Falck-Ytter\textsuperscript{a}, Holger Schünemann\textsuperscript{b}, Gordon Guyatt\textsuperscript{b,*}

\textsuperscript{a}Division of Gastroenterology, Case and VA Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA
\textsuperscript{b}Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

Accepted 17 January 2010

Authorities acknowledge that systematic reviews provide the optimal basis for collecting and assessing the evidence that bears on patient management recommendations. In his article introducing JCE’s series describing the Agency for Healthcare Research and Quality (AHRQ)’s effective health care program, Mark Helfand distinguishes between systematic reviews and “complex evidence reports” that address a broader range of questions, including “definition, diagnosis, management, and follow-up of a disease or condition.” Aside from definition, all these questions appear to us as an examination of alternative approaches to managing patients. Such issues are best addressed by structured questions and, if brought together in a single document, constitute a series of related systematic reviews or overviews of systematic reviews.

In this commentary, we address the conduct of such systematic reviews. We present a perspective arising from our participation in the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group. GRADE, a strategy for rating quality of evidence and grading strength of recommendations\textsuperscript{[1,2]}, has been endorsed by a wide variety of prominent organizations (American College of Physicians, World Health Organization, Cochrane Collaboration, National Institute for Clinical Excellence, and UpToDate) and is emerging as the dominant system for rating quality of evidence and strength of recommendations in guideline groups around the world.

In the current US environment, systematic reviews addressing alternative management strategies under the rubric of “comparative effectiveness reviews” have gained a high profile. Central to AHRQ’s evidence-based practice centers’ (EPCs)—and anyone else’s—effort to produce such reviews is the question of how to best conduct a thorough assessment of a body of evidence. Crucial to that issue is how one defines what we (and GRADE) call “quality” and what, in their article in JCE’s series describing EPC’s approach to comparative effectiveness reviews, Owen and authors call “strength of evidence.”

Although the forefathers of grading systems included aspects beyond “risk of bias”\textsuperscript{[3]}, quality has often been used to describe “risk of bias.” Although they do not explicitly define “strength of evidence,” Owens et al’s characterization is completely consistent with GRADE’s definition of quality of evidence from systematic reviews: the extent to which we can be confident in estimates of the magnitude of effect. As Owens and colleagues point out, quality is, therefore, much more than risk of bias, but includes domains, such as precision, consistency, and directness, as well as considerations of publication bias, magnitude of effect, and dose—response relationship.

Owens and colleagues describe the EPC approach as based in large measure on the GRADE working group approach. We agree, and in the remainder of this commentary, we will reflect on the possible differences between the GRADE and EPC approaches, differences that we view as minor.

Both systems use four categories of quality, three of which carry the same labels (high, moderate, and low) and share the same definitions; the final category is characterized by GRADE as “very low” and by the EPCs as “insufficient.” The term “insufficient,” as we understand it, implies insufficient to make a decision. This judgment is in the domain of a guideline panel rather than systematic review authors. Furthermore, the necessity to make decisions even when evidence is low quality—acknowledged explicitly by Owens and colleagues—may apply with equal force when evidence is of very low quality.

GRADE explicitly designates that randomized trials begin as high-quality evidence and may be rated down by limitations in each major area (risk of bias, imprecision, inconsistency, indirectness, and publication bias). Observational studies begin as low-quality evidence and may be rated up by a large magnitude of effect, a dose—response relationship, and an inference that plausible sources of bias could only diminish apparent effects or increase absent effects. This hierarchy of study designs...
underlies EPC judgments but is not made explicit in the same way.

Owens and colleagues include along with patient-important outcomes, surrogate markers among the major outcomes for their reviews. Although this statement is open to interpretation, it reflects a possible difference in approach. GRADE advocates focusing exclusively on patient-important outcomes and, when it is necessary to consider surrogates, to view surrogates only as indirect evidence for patient-important outcomes. This implies a need to estimate the magnitude of effect on patient-important outcomes and consider the uncertainty of relying on surrogates to estimate magnitude of effect. This is, however, completely consistent with the example that Owens and colleagues provide of indirectness introduced by the measurement of nutritional variables rather than the patient-important outcome of wound healing.

Owens and colleagues define consistency as “the degree to which reported effect sizes from included studies appear to have the same direction of effect.” In a comparison of intervention A and B, small, apparent effects that favor A may be completely consistent (ie, easily explained by chance) with small apparent effect that favors B. Thus, in considering consistency, GRADE focuses primarily on the magnitude of differences in estimates of effect and the associated precision of those estimates.

Another apparent difference we view as purely semantic. When the patients under consideration differ in important ways from those studied or the interventions under consideration differ from those studied, GRADE classifies the evidence as indirect. The EPCs classify such considerations under the rubric “applicability.” The underlying conceptual issues are, however, identical.

Determining the final quality of evidence for each important outcome requires careful consideration of all domains. Owens and colleagues acknowledge the strengths, in terms of transparency, of explicit decisions to rate up or down the quality of evidence and thus provide EPCs with the option of using GRADE’s approach in this regard (Fig. 1). They correctly point out the lack of evidence to choose between this approach and what they call a “qualitative approach”—which makes the same judgments without specific attribution to individual domains. Owens’ and colleagues’ wise counsel to EPCs to make the rationale for their decisions about final ratings of evidence quality explicit suggests that in the end the approaches differ very little.

We reiterate that these differences are of little importance. This is good news for the EPCs target audiences: they can interpret EPC’s strength of evidence ratings in the same way as the quality of evidence ratings used by the over 40 systematic review, guideline, and health technology assessment agencies worldwide that have adopted GRADE.

**References**

