Since 2004, the 14 Evidence-based Practice Centers (EPCs) in the United States and Canada have prepared comparative effectiveness reviews (CERs) for the Agency for Healthcare Research and Quality’s (AHRQ) Effective Health Care program [1]. From its inception, the Effective Health Care program has invited the public to comment on draft CERs. The articles in this issue of the Journal of Clinical Epidemiology address some of the most frequently mentioned methodological inconsistencies encountered in the first year and a half of public and peer comment on these reviews.

Most of the comments came from scientists within industry and professional groups. Other comments came to the program from stakeholders, either through a panel with representatives from insurance companies, industry, consumer groups, and other stakeholders that was convened in 2004 and met quarterly to discuss and comment on the program or through other channels.

Critics pointed out inconsistencies among reports in decisions that ranged from the way questions were framed to selection of articles for inclusion, inclusion of observational studies, techniques for pooling studies, and presentation and wording of comments on the overall strength of evidence. The reviewers also asked for more specific, customized methods for reviews for diagnostics, devices, and other categories of interventions.

For example, public comments across many reports showed that many considered the widespread practice of using a fixed effects model for harms [2] to represent a bias on the part of reviewers rather than a rational, fair, or consistent methodology. In one of the first CERs, a review of erythropoietin and darbepoetin for treating anemia in patients undergoing chemotherapy for cancer, the investigators used a fixed effects model to derive pooled estimates for mortality and adverse effects from trials of patients who had a wide variety of cancers [3]. A critique from a manufacturer argued that the studies were too clinically heterogeneous, and that the fixed effects approach was more likely than a random effects model to find a significant increase in mortality with treatment. It also pointed out that, in the absence of clear guidance, reviewer bias could affect the choice of statistical model.

The critics were right to say there was no programmatic guidance for how to handle these issues and that other EPCs might have handled them differently. As discussed in the article by Chou et al. [4], however, there is merit in using a broadly inclusive meta-analysis to study a drug’s adverse effects. In some circumstances, there is also merit in using a fixed effects statistical model, particularly when evaluating a rare harm. However, some statisticians would not use a fixed effects model unless that the adverse outcome of interest was relatively rare and idiosyncratic, whereas in a body of studies of cancer treatments, death is a common outcome.

The role of observational studies in CERs was another major theme. Like Chou et al. [4], Vandenbergroucke [5–7] strongly endorses the principle that observational studies should be considered routinely in evaluating harms. The EPC guide recognizes that many types of studies are important. Helfand and Balshem [8], in the section of their article about “Types of Evidence,” describe a hierarchy in which observational studies that reflect real-world effectiveness [9,10] and have other characteristics of comparative effectiveness research rank higher than certain randomized trials. A separate article in this series on the role of observational studies in assessing benefits is under review at this journal.

The Effective Health Care program also solicits topic nominations from the public. In her article, Evelyn Whitlock et al. [11] describe a process for preparing briefing documents and the criteria the program uses to prioritize nominations. Returning to nominators or other stakeholders to get more information about the context, scope, and need for a review is an important aspect of this work. The Institute of Medicine panel on priorities for comparative effectiveness research has already cited this work, endorsing the use of topic briefing documents, the application of preset criteria, and other aspects of this process [12].

The last article in this group concerns grading a body of evidence [13]. The guidance in this article was developed by a work group consisting of EPC investigators from the United States and Canada, many of whom had participated in developing or evaluating some of the most widely used systems for grading evidence [14–16].

While recognizing the important contributions of the US Preventive Services Task Force (USPSTF) and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group in this area, the work group was influenced by three factors to adapt these
systems to the particular needs of the Effective Health Care program.

The first is that CERs present unique challenges [8]. The scope and purpose of these reviews influence the logistics of conducting and reporting the review, straining the credibility of conventional strategies for searching and selecting articles, types of articles that are the most relevant and methodologically sound, statistical methods for combining studies, and, in particular, methods for summarizing or grading the evidence about multiple comparisons among competing interventions.

Current practice in AHRQ evidence reports emerged from two related but different traditions, which for convenience can be called “systematic reviews” and “complex evidence reports.” Cochrane reviews best exemplify the “systematic review” tradition. Typically, these reviews focus on a narrow question or set of questions—“is atenolol superior to placebo for heart failure?”—and, until recently, most of them focused on meta-analysis of a group of small, similar explanatory randomized trials. CERs are better described as “complex evidence reports” than as “systematic reviews.” Although they are systematic, CERs [8,10,11] emerge from the fields of medical decision making and technology assessment, as well as clinical epidemiology. They may address a series of questions related to the definition, diagnosis, management, and follow-up of a disease or condition and usually encompass up to several dozen questions about a group of interventions. Historically, they encompass a broad range of research designs and practical issues in application and implementation in addition to efficacy [17].

Focus on a narrow question—“is atenolol superior to placebo for heart failure?”—permits a greater degree of standardization of methods but at the cost of relevance to decision makers. Certainly there is an audience for such a review—for example, individuals who have been prescribed atenolol for heart failure would certainly be interested in the results. But for most patients with heart failure and the clinicians who see them, a broader set of questions—which beta blocker should I take or prescribe, for whom, and when?—is more pertinent. And some decision makers, such as a guideline or formulary committee, require an even broader scope, comparing most or all available beta blockers for most indications, from hypertension to heart failure, migraine, and esophageal varices to decide which ones are essential or uniquely useful for at least some patients. Many attempts to use the existing grading systems for a review with such a scope have had disappointing results.

A second factor was the need to place particular emphasis on considering applicability in grading. The USPSTF system and GRADE are designed for use by a defined group of decision makers, usually, guideline developers. The AHRQ CERs are intended for several audiences, not all of whom will agree about to whom, or to what circumstances, the evidence should be applicable.

These two factors had a strong influence on how the group adapted GRADE. Both the EPCs and GRADE use the term “directness” to evaluate the relation between intermediate outcomes and patient-important health outcomes and for indirect comparisons of treatments. In GRADE, however, directness also encompasses applicability, whereas, in the EPC system, reviewers grade the applicability of a body of evidence separately. The system for assessing applicability will be described in a separate article, which is under review at this journal.

The third factor was the group’s own experience in working with guidelines, coverage decisions, and others who use evidence reviews to make decisions or improve care. The article introducing this series refers to the importance of identifying clinical theories and the causal logic underlying the opinions of those who believe that a particular intervention is superior to the alternatives [8]. As discussed in that article, recognizing the importance of clinical and biomedical beliefs in decision making has been a hallmark of the EPC program since its inception in 1997. In many reviews, particularly those conducted for the USPSTF and the US Medicare (MEDca) program, EPCs have used an analytic framework or causal pathway to organize the presentation of the results of a review. The EPC grading work group sought to highlight this approach to synthesizing and summarizing a body of evidence.

Involvement of the public made the need for better standardization and justification for CER methods clear to the EPCs and to their sponsor, the AHRQ [1].

Despite the lack of definitive, final answers to the issues it addresses, the guidance in these articles can improve the credibility and consistency of CERs. The central idea behind systematic review is to avoid bias in identifying, selecting, or interpreting scientific literature. From the viewpoint of pharmaceutical, device manufacturers, and physicians whose products or practices are under study, a higher degree of standardization provides a more predictable process and a more “level playing field” in which the way evidence is evaluated and synthesized is less dependent on the practice styles of different reviewers.


