EDITORIAL

STEVEN WOLOSHIN, MD, MS
Associate Editor

Eff Clin Pract. 2000;3:94–95.

Arguing about Grades

As long as there have been grades, people have been arguing about them. It is no different in this issue of **ecp**, in which Roman and colleagues[1] criticize the standard approach to grading the evidence for clinical practices—the hierarchical grading system (i.e., grade I, II, III) popularized by the U.S. Preventive Services Task Force.

Their main criticism is that the hierarchical grading system is too inflexible to accurately characterize the evidence for many practices in diabetes care. Consider, for example, the evidence for hemoglobin A_{1c} monitoring. Because this procedure was only one part of the Diabetes Control and Complications Trial,[2] the author's argue that it would currently be accorded a grade of III (i.e., expert opinion). In this way, the hierarchical system tends to undervalue good evidence if it does not precisely fit one of very few narrow categories. Many reasonable practices may therefore be excluded from "evidence-based" practice guidelines. Ultimately, the hierarchical grading system may, if strictly interpreted, limit the realm of evidence-based medicine to the few practices that have been explicitly validated in randomized trials that enrolled the specific population of interest.

In place of the traditional hierarchical system, Roman and colleagues propose a descriptive grading system. The main advantages of the new system are its precision and clarity. Practices generalized or adapted from randomized trials, for example, would not be demoted to grade III evidence but would be designated as what they are: an embedded component of a trial, a trial-based practice applied to a new population (or subpopulation), and so on. Moreover, the new grades are simpler to interpret. The opaque grades I, II-1, II-2, II-3, and III would be replaced by such self-defining terms as "RCT–embedded component" or "RCT–different population."

Why should you care about another argument over grades? (Didn't you get your fill in high school?) Physicians should care because grading systems influence which practices they are expected to perform. Furthermore, physicians will increasingly find themselves graded on how well they follow practice guidelines. Policymakers should care because they need reliable, usable grades to help them decide which guidelines to adopt. Roman and colleagues improve on the current grading system by being more explicit about what the evidence is. They acknowledge, however, that important challenges remain. Notably, grades need to account not only for study design but also for study quality. For example, it does not make sense to give the same grade to a large multicenter randomized trial with 40,000 participants and to an unblinded randomized trial with 100 participants.

But there is one other issue that the authors did not address: grade inflation, which can create the impression that the evidence for a given practice is stronger than it really is. Where the traditional grading system tends to undervalue evidence (e.g., hemoglobin A_{1c} testing outside a treatment program like the Diabetes Control and Complications Trial would be downgraded to "expert opinion"), the descriptive system may do the opposite (e.g., it is tempting to read a grade such as "RCT–component" as "RCT"). The problem with grade inflation is that it may encourage too much extrapolation from the direct evidence.

Recent history warns us that it can be dangerous to assume that evidence of benefit in one context can be safely extrapolated to another. For example, it might be tempting to extrapolate from the successful treatment of symptomatic post–myocardial infarction arrhythmias to treating asymptomatic arrhythmias as well, or to generalize from the effectiveness of early cardiac catheterization for patients who have had myocardial infarction and ongoing ischemia and advocate early invasive management

This paper is available at ecp.acponline.org.

for all patients with non–Q-wave infarctions. When specifically tested in randomized trials, both of the foregoing extrapolations have in fact been associated with higher patient mortality.[3, 4] In short, grade inflation may do more harm than good. Addressing when and how much extrapolation is reasonable represents a huge challenge to any grading system.

We are interested in your thoughts on how to do a better job grading the evidence. Send them to us by e-mail at ecp@mac.dartmouth.edu. We'll be happy to give you extra credit.

References

1. Roman SH, Silberzweig SB, Siu AL. Grading the evidence for diabetes performance measures. Eff Clin Pract. 2000;3:85-91.
2. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial Research Group. N Engl J Med. 1993; 329:977-86.
3. Boden WE, O'Rourke RA, Crawford MH, et al. Outcomes in patients with acute non-Q-wave myocardial infarction randomly assigned to an invasive as compared with a conservative management strategy. Veterans Affairs Non-Q-Wave Infarction Strategies in Hospital (VANQWISH) Trial Investigators. N Engl J Med. 1998;338:1785-92.
4. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. N Engl J Med. 1989; 321:406-12.