CHEST Special Features

Grading Strength of Recommendations and Quality of Evidence in Clinical Guidelines*

Report From an American College of Chest Physicians Task Force

Gordon Guyatt, MD, MSc, FCCP; David Gutterman, MD, FCCP; Michael H. Baumann, MD, MSc, FCCP; Doreen Addrizzo-Harris, MD, FCCP; Elaine M. Hylek, MD, MPH; Barbara Phillips, MD, FCCP; Gary Raskob, PhD; Sandra Zelman Lewis, PhD; and Holger Schünemann, MD, PhD, FCCP

While grading the strength of recommendations and the quality of underlying evidence enhances the usefulness of clinical guidelines, the profusion of guideline grading systems undermines the value of the grading exercise. An American College of Chest Physicians (ACCP) task force formulated the criteria for a grading system to be utilized in all ACCP guidelines that included simplicity and transparency, explicitness of methodology, and consistency with current methodological approaches to the grading process. The working group examined currently available systems, and ultimately modified an approach formulated by the international GRADE group. The grading scheme classifies recommendations as strong (grade 1) or weak (grade 2), according to the balance among benefits, risks, burdens, and possibly cost, and the degree of confidence in estimates of benefits, risks, and burdens. The system classifies quality of evidence as high (grade A), moderate (grade B), or low (grade C) according to factors that include the study design, the consistency of the results, and the directness of the evidence. For all future ACCP guidelines, The College has adopted a simple, transparent approach to grading recommendations that is consistent with current developments in the field. The trend toward uniformity of approaches to grading will enhance the usefulness of practice guidelines for clinicians. ***(CHEST 2006; 129:174–181)***

Key words: grading recommendations; grading system; methodology

Abbreviations: ACCP = American College of Chest Physicians; RCT = randomized controlled trial; RRR = relative risk reduction

Treatment decisions involve a tradeoff between benefits on the one hand, and risks, burdens, and, potentially, costs on the other. Guideline panels provide recommendations for the management of typical patients. To integrate these recommendations with their own clinical judgment, and with individual patient values and preferences, clinicians need to understand the basis for the recommendations that expert guidelines offer. A systematic approach to grading the strength of management recommendations can minimize bias and aid interpretation.[3] Indeed, most guideline groups have accepted the necessity for some sort of grading scheme.

While the grading of recommendations represents

*From the Departments of Medicine, and Clinical Epidemiology and Biostatistics (Dr. Guyatt), McMaster University, Hamilton, ON, Canada; the Department of Medicine (Dr. Gutterman), Medical College of Wisconsin, Milwaukee, WI; the University of Mississippi Medical Center (Dr. Baumann), Jackson, MS; New York University School of Medicine (Dr. Addrizzo-Harris), New York, NY; the Department of Medicine (Dr. Hylek), Research Unit-Section of General Internal Medicine, Boston University School of Medicine, Boston, MA; the University of Kentucky College of Medicine (Dr. Phillips), Lexington, KY; College of Public Health (Dr. Raskob), University of Oklahoma Health Sciences Center, Oklahoma City, OK; the American College of Chest Physicians (Dr. Lewis), Northbrook, IL; and the Department of Clinical Epidemiology, Italian National Cancer Institute (Dr. Schünemann), Rome, Italy.

Manuscript received August 16, 2005; revision accepted August 21, 2005.

Correspondence to: Gordon Guyatt, MD, MSc, FCCP, Department of Clinical Epidemiology and Biostatistics, HSC-2C12, McMaster University, 1200 Main St West, Hamilton, ON, Canada L8N 3Z5; e-mail: guyatt@mcmaster.ca

a positive development for guideline development and interpretation, the proliferation of grading systems has proved to be an unfortunate consequence. Methodologists and guideline developers have given much thought and effort to considering the criteria and approaches to an optimal grading system. The American College of Chest Physicians (ACCP) convened a working group to review the issue and to agree on a grading system that would be consistent with the latest developments in the field.

For editorial comment see pages 7 and 10

The task force began by developing criteria that define an optimal grading system (Table 1), placing them in an order that approximates their relative importance. These criteria guided the decisions of the group in the choice of the grading system that follows.

Strength of Recommendation

Guideline panels should make recommendations to administer, or not administer, an intervention, on the basis of tradeoffs between benefits on the one hand, and risks, burdens, and, potentially, costs on the other. If benefits outweigh risks and burdens, experts will recommend that clinicians offer a treatment to appropriately chosen patients. The uncertainty associated with the tradeoff between the benefits and the risks and burdens will determine the strength of recommendations.

The ACCP task force chose to classify recommendations into two levels, strong and weak (Table 2). If guideline panelists are very certain that benefits do, or do not, outweigh risks and burdens, they will make a strong recommendation, grade 1. If they think that the benefits and the risks and burdens are finely balanced, or if appreciable uncertainty exists about the magnitude of the benefits and risks, they must offer a weak, grade 2 recommendation.

A two-level grading system has the merit of simplicity. Two levels also facilitate the clear interpretation of the implications of strong and weak recommendations by clinicians. We offer three ways that clinicians can interpret strong and weak recommendations. We have already presented the first way. A strong recommendation signifies that benefits clearly outweigh the risks, or the reverse; a weak recommendation signifies that benefits and risks are closely balanced, or uncertain.

Clinicians are becoming increasingly aware of the importance of patient values and preferences in clinical decision making.[4] A second way to interpret strong and weak recommendations is in relation to patient values and preferences. For decisions in which it is clear that benefits far outweigh risks, or risks far outweigh benefits, virtually all patients will make the same choice (see box 1 for an example). In such instances, guideline panels can offer a strong (grade 1) recommendation. In contrast, there are other choices in which patient values and preferences will play a crucial role and in which patients will, as a result, make different choices. See boxes 2

Box 1: Short-term aspirin reduces the relative risk of death after myocardial infarction by approximately 25%. Aspirin has minimal side effects and very low cost. Peoples' values and preferences are such that virtually all patients suffering a myocardial infarction would, if they understood the choice they were making, opt to receive aspirin. Guideline panels can thus offer a strong recommendation for aspirin administration in this setting.

Box 2: Consider a patient a 40 year-old man who has suffered an idiopathic deep venous thrombosis and has been taking adjusted dose warfarin for one year. If the patient continues on standard-intensity warfarin his risk of recurrent DVT will be reduced by approximately 10% per year.[1] The inevitable burdens of the treatment include taking a warfarin pill daily, keeping dietary intake of vitamin K constant, monitoring the intensity of anticoagulation with blood tests, and living with the increased risk of both minor and major bleeding. Some patients who are very averse to a recurrent DVT may consider the down sides of taking warfarin well worth it. Others are likely to consider the benefit not worth the risks and inconvenience.

Table 1—*Criteria for an Optimal Grading System*

Criteria	Description
1	Separation of grades of recommendations from quality of evidence
2	Simplicity and transparency for clinician consumer
3	Sufficient (but not too many) categories
4	Explicitness of methodology for guideline developers
5	Simplicity for guideline developers
6	Consistent with general trends in grading systems
7	Explicit approach to different levels of evidence for different outcomes