RATING QUALITY OF EVIDENCE AND STRENGTH OF RECOMMENDATIONS

GRADE: going from evidence to recommendations

The GRADE system classifies recommendations made in guidelines as either strong or weak. This article explores the meaning of these descriptions and their implications for patients, clinicians, and policy makers

This is the third of a series of five articles describing the GRADE approach to developing and presenting recommendations for management of patients. In it, we deal with how GRADE suggests clinicians should interpret the strength of a recommendation.

What do we mean by the strength of a recommendation?

The strength of a recommendation reflects the extent to which we can be confident that the desirable effects of an intervention outweigh the undesirable effects. Desirable effects of an intervention include reduction in morbidity and mortality, improvement in quality of life, reduction in the burden of treatment (such as having to take drugs or the inconvenience of blood tests), and reduced resource expenditures. Undesirable consequences include adverse effects that have a deleterious impact on morbidity, mortality, or quality of life or increase use of resources.

Previous grading systems have used up to nine categories of strength of recommendations.[1] The GRADE system has only two categories—although in this article we will characterise them as strong and weak, guideline panels may choose different words to characterise the two categories of strength. When using GRADE, panels make strong recommendations when they are confident that the desirable effects of adherence to a recommendation outweigh the undesirable effects. Weak recommendations indicate that the desirable effects of adherence to a recommendation probably outweigh the undesirable effects, but the panel is less confident.

Strong and weak recommendations provide specific guidance

GRADE's binary classification of strength of recommendations provides clear direction to patients, clinicians, and policy makers. The implications of a strong recommendation are:

- For patients—most people in your situation would want the recommended course of action and only a small proportion would not; request discussion if the intervention is not offered
- For clinicians—most patients should receive the recommended course of action
- For policy makers—the recommendation can be adopted as a policy in most situations.

The implications of a weak recommendation are:

- For patients—most people in your situation would want the recommended course of action, but many would not
- For clinicians—you should recognise that different choices will be appropriate for different patients and that you must help each patient to arrive at a management decision consistent with her or his values and preferences
- For policy makers—policy making will require substantial debate and involvement of many stakeholders.

As clinicians become more aware of variability in patients' values and preferences, they are turning to structured decision aids to facilitate the decision making process.[2] A strong recommendation indicates that use of a decision aid is unnecessary—almost all informed patients will make the same choice. A weak recommendation indicates that a decision aid could be useful.

Managers of healthcare systems are becoming increasingly interested in ensuring the quality of care. Guidelines help managers to differentiate practices that constitute quality of care from others that are discretionary. GRADE provides clear guidance on these matters: The management options associated with strong, but not with weak, recommendations are candidates for quality criteria. When a recommendation is weak, discussing with patients and families the relative merits of the alternative management strategies may become a quality criterion.

Four key factors determine the strength of a recommendation

The first key determinant of the strength of a recommendation is the balance between the desirable and undesirable consequences of the alternative management strategies, on the basis of the best estimates of those consequences (table 1). Consider, for instance, the use of antenatal steroids in women destined to deliver an infant prematurely. Administration of steroids to mothers decreases the risk of infant respiratory distress syndrome with minimal side effects, inconvenience, and costs. Advantages of steroid administration hugely outweigh the disadvantages, indicating the appropriateness of a strong recommendation.

When advantages and disadvantages are closely

Gordon H Guyatt professor, CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada L8N 3Z5

Andrew D Oxman researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs Plass, 0130 Oslo, Norway

Regina Kunz associate professor, Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, 4031 Basel, Switzerland

Yngve Falck-Ytter assistant professor, Division of Gastroenterology, Case Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA

Gunn E Vist researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs Plass, 0130 Oslo, Norway

Alessandro Liberati associate professor, University of Modena and Reggio Emilia and Agenzia Sanitaria Regionale, Bologna, Italy

Holger J Schünemann associate professor, Department of Epidemiology, CLARITY Research Group, Italian National Cancer Institute Regina Elena, Rome, Italy

For the GRADE Working Group

Correspondence to: guyatt@mcmaster.ca

This is the third in a series of five articles that explain the GRADE system for rating the quality of evidence and strength of recommendations

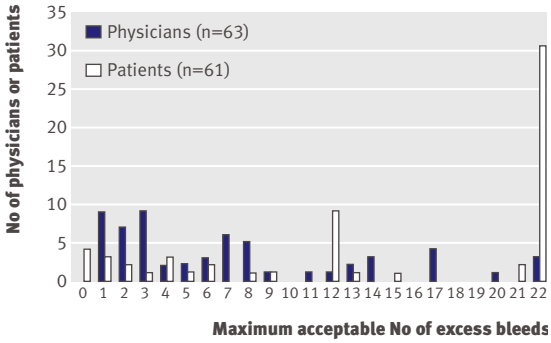

Fig 1 Varying thresholds of major gastrointestinal bleeding found acceptable by patients and physicians for prevention of eight strokes in 100 patients

balanced, a weak recommendation becomes appropriate. Consider, for instance, patients with atrial fibrillation at low risk of stroke. Warfarin can reduce that low risk even further, but adds inconvenience and an increased risk of bleeding. The right choice under such circumstances is likely to differ between patients.

The second determinant of the strength of a recommendation is the quality of the evidence. If we are uncertain of the magnitude of the benefits and harms of an intervention, making a strong recommendation for or against a particular course of action becomes problematic. For instance, graduated compression stockings have an apparent large effect in reducing deep venous thrombosis in people making long plane journeys. The randomised trials from which the estimate of effect comes were, however, seriously flawed—the techniques for measuring deep venous thrombosis were not reproducible, and the studies were unblinded. Despite the apparent large benefit, use of stockings warrants only a weak recommendation.[3]

The third determinant of the strength of recommendation is uncertainty about, or variability in, values and preferences. Given that alternative management strategies will always have advantages and disadvantages, and thus a trade-off exists, how a guideline panel values benefits, risks, and inconvenience is critical to the strength of any recommendation.

Consider the subject of preventing strokes in patients with atrial fibrillation. Warfarin, relative to no antithrombotic therapy, reduces the risk of stroke by approximately 65% but increases the risk of severe gastrointestinal bleeding. Devereaux and colleagues asked 63 physicians and 61 patients how many serious gastrointestinal bleeds they would tolerate in 100 patients and still be willing to prescribe or take warfarin to prevent eight strokes (four minor and four major) in 100 patients.[4] Figure 1 shows the results. Whereas physicians gave a wide diversity of responses, most patients placed a high value on avoiding a stroke and were ready to accept a bleeding risk of 22% to reduce their chances of having a stroke by 8%. Even among patients, however, diversity in values and preferences was apparent; a few patients were ready to accept only a small risk of bleeding. These data suggest that only in patients at high risk of stroke would a strong recommendation for warfarin be warranted.

Contrast this with the decision faced by pregnant women with deep venous thrombosis. Warfarin treatment between the sixth and 12th week of pregnancy puts women's unborn infants at risk of relatively minor developmental abnormalities. The alternative, heparin, eliminates the risk to the child. The benefit, however, comes with disadvantages of pain, inconvenience, and cost. Clinicians' experience is that women overwhelmingly place a high value on preventing fetal complications. Thus, despite its disadvantages, a strong recommendation for heparin substitution is warranted.

The final determinant of the strength of a recommendation is cost. Cost is much more variable over time and geographical areas than are other outcomes. Drug costs tend to plummet when patents expire, and charges for the same drug differ widely across jurisdictions. In addition, the resource implications vary widely. For instance, a year's prescription of the same expensive drug may pay for a single nurse's salary in the United States and 30 nurses' salaries in China.

Thus, although higher costs reduce the likelihood of a strong recommendation in favour of an intervention, the context of the recommendation will be critical. In considering resource allocation, guideline panels must therefore be specific about the setting to which a recommendation applies.

Strong recommendations may not be important from all perspectives

If the consequences of the choice are relatively unimportant, some patients may not bother with even strong recommendations. This is particularly likely if they are faced with many new drugs or many suggestions to change their lifestyle.

When setting priorities, governments and public health officials must also consider factors beyond the strength of a recommendation. These include the prevalence of the health problem, considerations of equity, and the potential for improvement in quality of care, all of which will have an impact on the population health gain of an intervention.

Determinants of strength of recommendation

Factor	Comment
Balance between desirable and undesirable effects	The larger the difference between the desirable and undesirable effects, the higher the likelihood that a strong recommendation is warranted. The narrower the gradient, the higher the likelihood that a weak recommendation is warranted
Quality of evidence	The higher the quality of evidence, the higher the likelihood that a strong recommendation is warranted
Values and preferences	The more values and preferences vary, or the greater the uncertainty in values and preferences, the higher the likelihood that a weak recommendation is warranted
Costs (resource allocation)	The higher the costs of an intervention—that is, the greater the resources consumed—the lower the likelihood that a strong recommendation is warranted

Quality of evidence	
High quality	⊕⊕⊕⊕ or A
Moderate quality	⊕⊕⊕◯ or B
Low quality	⊕⊕◯◯ or C
Very low quality	⊕◯◯◯ or D
Strength of recommendation	
Strong recommendation for using an intervention	↑ ↑ or 1
Weak recommendation for using an intervention	↑ ? or 2
Weak recommendation against using an intervention	↓ ? or 2
Strong recommendation against using an intervention	↓ ↓ or 1

Fig 2 Representations of quality of evidence and strength of recommendations

Recommendations to use interventions in research context may be appropriate

Guideline panels may face decisions about promising interventions associated with appreciable harms or costs and with insufficient evidence of benefit to support their use. They may be reluctant to close the door on such an intervention or to inappropriately provide a weak recommendation for its use. Their fears will be realised if the appropriate recommendation against use of the intervention in clinical practice has the effect of stifling further investigation.

Recommendation for use of an intervention only in the context of research may ameliorate these problems. Furthermore, such a recommendation may encourage efforts to answer important research questions. The National Institute for Health and Clinical Excellence felt this was the case in eight of its first 95 technology appraisals, which included recommendations for use in the context of research.

Various presentations of quality of evidence and strength of recommendations may be appropriate

Most guideline panels have used letters and numbers to summarise their recommendations, but they have used them differently. This is potentially confusing.[5] Symbolic representations of quality of evidence and strength of recommendations are appealing in that they are free of this history. On the other hand, organisations may have good reasons for choosing letters and numbers. Clinicians seem to be very comfortable with numbers and letters, and these are particularly suitable for verbal communication.

GRADE offers preferred symbolic representations and, for organisations that wish to use numbers and letters, a preferred number/letter representation, for quality of evidence and grades of recommendation (fig 2).[5]

The members of the Grade Working Group are Phil Alderson, Pablo Alonso-Coello, Jeff Andrews, David Atkins, Hilda Bastian, Hans de Beer, Jan Brozek, Francoise Cluzeau, Jonathan Craig, Ben Djulbegovic, Yngve Falck-Ytter, Beatrice Fervers, Signe Flottorp, Paul Glasziou, Gordon H Guyatt, Robin Harbour, Margaret Haugh, Mark Helfand, Sue Hill, Roman Jaeschke, Katharine Jones, Ilkka Kunnamo, Regina Kunz, Alessandro Liberati, Nicola Magrini, Merce Marzo, James Mason, Jacek Mrukowics, Andrew D Oxman, Susan Norris, Vivian Robinson, Holger J Schünemann, Jane Thomas, Tessa Tan Torres, David Tovey, Peter Tugwell, Mariska Tuut, Helena Varonen, Gunn E Vist, Craig Wittington, John Williams, and James Woodcock.

Contributors: All listed authors, and other members of the GRADE working group, contributed to the development of the ideas in the manuscript and read and approved the manuscript. GHG wrote the first draft and collated comments from authors and reviewers for subsequent iterations. All other listed authors contributed ideas about structure and content, provided examples, and reviewed successive drafts of the manuscript and provided feedback. GHG is the guarantor.

Funding: None.

Competing interests: All authors are involved in the dissemination of GRADE, and GRADE's success has a positive influence on their academic careers. Authors listed on the byline have received travel reimbursement and honoraria for presentations that included a review of GRADE's approach to rating quality of evidence and grading recommendations. GHG acts as a consultant to UpToDate; his work includes helping UpToDate in their use of GRADE. HJS is documents editor and methodologist for the American Thoracic Society; one of his roles in these positions is helping implement the use of GRADE. HJS is supported by "The human factor, mobility and Marie Curie Actions Scientist Reintegration European Commission Grant: IGR 42192—GRADE." AL is helping the use of GRADE by different institutions in the Italian health service, and in this role he has implemented GRADE to produce clinical recommendations in oncology through Grant No 249 (2005-7), Bando Ricerca Finalizzata, Ministero della Salute, Roma, Italy.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Fleisher LA, Bass EB, McKeown P. Methodological approach: American College of Chest Physicians guidelines for the prevention and management of postoperative atrial fibrillation after cardiac surgery. *Chest* 2005;128:17-23S.
2 O'Connor AM, Stacey D, Entwistle V, Llewellyn-Thomas H, Rovner D, Holmes-Rovner M, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev* 2003;(1):CD001431.
3 Geerts W, Ray JG, Colwell CW, Bergqvist D, Pineo GF, Lassen MR, et al. Prevention of venous thromboembolism. *Chest* 2005;128:3775-6.
4 Devereaux PJ, Anderson DR, Gardner MJ, Putnam W, Flowerdew GJ, Brownell BF, et al. Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study. *BMJ* 2001;323:1218-22.
5 Schunemann HJ, Best D, Vist G, Oxman AD. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. *CMAJ* 2003;169:677-80.

SUMMARY POINTS

The strength of a recommendation reflects the extent to which we can be confident that desirable effects of an intervention outweigh undesirable effects

GRADE classifies recommendations as strong or weak

Strong recommendations mean that most informed patients would choose the recommended management and that clinicians can structure their interactions with patients accordingly

Weak recommendations mean that patients' choices will vary according to their values and preferences, and clinicians must ensure that patients' care is in keeping with their values and preferences

Strength of recommendation is determined by the balance between desirable and undesirable consequences of alternative management strategies, quality of evidence, variability in values and preferences, and resource use

Endpiece

Two kinds of surgery

Surgical operations are of two kinds—those that benefit the patient and those that kill him.

Abu al-Qasim Khalaf bin 'Abbas el-Zahrawi, also known as Albucasis (940-1013)

Submitted by **Munier Hossain**, staff grade orthopaedic surgeon, Ysbyty Gwynedd, Bangor