RATING QUALITY OF EVIDENCE AND STRENGTH OF RECOMMENDATIONS

GRADE: an emerging consensus on rating quality of evidence and strength of recommendations

Guidelines are inconsistent in how they rate the quality of evidence and the strength of recommendations. This article explores the advantages of the GRADE system, which is increasingly being adopted by organisations worldwide

Guideline developers around the world are inconsistent in how they rate quality of evidence and grade strength of recommendations. As a result, guideline users face challenges in understanding the messages that grading systems try to communicate. Since 2006 the *BMJ* has requested in its "Instructions to Authors" on bmj.com that authors should preferably use the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system for grading evidence when submitting a clinical guidelines article. What was behind this decision?

In this first in a series of five articles we will explain why many organisations use formal systems to grade evidence and recommendations and why this is important for clinicians; we will focus on the GRADE approach to recommendations. In the next two articles we will examine how the GRADE system categorises quality of evidence and strength of recommendations. The final two articles will focus on recommendations for diagnostic tests and GRADE's framework for tackling the impact of interventions on use of resources.

GRADE has advantages over previous rating systems (box 1). Other systems share some of these advantages, but none, other than GRADE, combines them all.[1]

Box 1 | Advantages of GRADE over other systems

- Developed by a widely representative group of international guideline developers
- Clear separation between quality of evidence and strength of recommendations
- Explicit evaluation of the importance of outcomes of alternative management strategies
- Explicit, comprehensive criteria for downgrading and upgrading quality of evidence ratings
- Transparent process of moving from evidence to recommendations
- Explicit acknowledgment of values and preferences
- Clear, pragmatic interpretation of strong versus weak recommendations for clinicians, patients, and policy makers
- Useful for systematic reviews and health technology assessments, as well as guidelines

What is "quality of evidence" and why is it important?

In making healthcare management decisions, patients and clinicians must weigh up the benefits and downsides of alternative strategies. Decision makers will be influenced not only by the best estimates of the expected advantages and disadvantages but also by their confidence in these estimates. The cartoon depicting the weather forecaster's uncertainty captures the difference between an assessment of the likelihood of an outcome and the confidence in that assessment (figure). The usefulness of an estimate of the magnitude of intervention effects depends on our confidence in that estimate.

Expert clinicians and organisations offering recommendations to the clinical community have often erred as a result of not taking sufficient account of the quality of evidence.[2] For a decade, organisations recommended that clinicians encourage postmenopausal women to use hormone replacement therapy.[3] Many primary care physicians dutifully applied this advice in their practices.

A belief that such therapy substantially decreased women's cardiovascular risk drove this recommendation. Had a rigorous system of rating the quality of evidence been applied at the time, it would have shown that because the data came from observational studies with inconsistent results, the evidence for a reduction in cardiovascular risk was of very low quality.[4] Recognition of the limitations of the evidence would have tempered the recommendations. Ultimately, randomised controlled trials have shown that hormone replacement therapy fails to reduce cardiovascular risk and may even increase it.[5 6]

The US Food and Drug Administration licensed the antiarrhythmic agents encainide and flecainide for use in patients on the basis of the drugs' ability to reduce asymptomatic ventricular arrhythmias associated with sudden death. This decision failed to acknowledge that because arrhythmia reduction reflected only indirectly on the outcome of sudden death the quality of the evidence for the drugs' benefit was of low quality. Subsequently, a randomised controlled trial showed that the two drugs increase the risk of sudden death.[7] Appropriate attention to the low quality of the evidence would have saved thousands of lives.

Failure to recognise high quality evidence can cause similar problems. For instance, expert recommendations lagged a decade behind the evidence from well conducted randomised controlled trials that thrombolytic therapy achieved a reduction in mortality in myocardial infarction.[8]

Insufficient attention to quality of evidence risks inappropriate guidelines and recommendations that may lead clinicians to act to the detriment of their

Gordon H Guyatt professor, 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

Gunn E Vist 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

Pablo Alonso-Coello researcher, Iberoamerican Cochrane Center, Servicio de Epidemiología Clínica y Salud Pública (Universidad Autónoma de Barcelona), Hospital de Sant Pau, Barcelona 08041, Spain

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

for the GRADE Working Group

Correspondence to:
G H Guyatt, CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, Room 2C12, 1200 Main Street, West Hamilton, ON, Canada L8N 3Z5
guyatt@mcmaster.ca

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

"I figure there's a 40% chance of showers, and a 10% chance we know what we're talking about."

patients. Recognising the quality of evidence will help to prevent these errors.

How should guideline developers alert clinicians to quality of evidence?

A formal system that categorises quality of evidence—for example, from high to very low—represents an obvious strategy for conveying quality of evidence to clinicians. Some limitations, however, do exist. Quality of evidence is a continuum; any discrete categorisation involves some degree of arbitrariness. Nevertheless, advantages of simplicity, transparency, and vividness outweigh these limitations.

What is "strength of recommendation" and why is it important?

A recommendation to offer patients a particular treatment may arise from large, rigorous randomised controlled trials that show consistent impressive benefits with few side effects and minimal inconvenience and cost. Such is the case with using a short course of oral steroids in patients with exacerbations of asthma. Clinicians can offer such treatments to almost all their patients with little or no hesitation.

Alternatively, treatment recommendations may arise from observational studies and may involve appreciable harms, burdens, or costs. Deciding whether to use antithrombotic therapy in pregnant women with prosthetic heart valves involves weighing the magnitude of reduction in valve thrombosis against inconvenience, cost, and risk of teratogenesis. Clinicians offering such treatments must help patients to weigh up the desirable and undesirable effects carefully according to their values and preferences.

Guidelines and recommendations must therefore indicate whether *(a)* the evidence is high quality and the desirable effects clearly outweigh the undesirable effects, or *(b)* there is a close or uncertain balance. A simple, transparent grading of the recommendation can effectively convey this key information.

There are limitations to formal grading of recommendations. Like the quality of evidence, the balance between desirable and undesirable effects reflects a continuum. Some arbitrariness will therefore be associated with placing particular recommendations in categories such as "strong" and "weak." Most organisations producing guidelines have decided that the merits of an explicit grade of recommendation outweigh the disadvantages.

What makes a good grading system?

Not all grading systems separate decisions regarding the quality of evidence from strength of recommendations. Those that fail to do so create confusion. High quality evidence doesn't necessarily imply strong recommendations, and strong recommendations can arise from low quality evidence.

For example, patients who experience a first deep venous thrombosis with no obvious provoking factor must, after the first months of anticoagulation, decide whether to continue taking warfarin long term. High quality randomised controlled trials show that continuing warfarin will decrease the risk of recurrent thrombosis but at the cost of increased risk of bleeding and inconvenience. Because patients with varying values and preferences will make different choices, guideline panels addressing whether patients should continue or terminate warfarin should—despite the high quality evidence—offer a weak recommendation.

Consider the decision to administer aspirin or paracetamol (acetaminophen) to children with chicken pox. Observational studies have observed an association between aspirin administration and Reye's syndrome.[9] Because aspirin and paracetamol are similar in their analgesic and antipyretic effects, the low quality evidence regarding the association between aspirin and Reye's syndrome does not preclude a strong recommendation for paracetamol.

Systems that classify "expert opinion" as a category of evidence also create confusion. Judgment is necessary for interpretation of all evidence, whether that evidence is high or low quality. Expert reports of their clinical experience should be explicitly labelled as very low quality evidence, along with case reports and other uncontrolled clinical observations.

Grading systems that are simple with respect to judgments both about the quality of the evidence and the strength of recommendations facilitate use by patients, clinicians, and policy makers.[1] Detailed and explicit criteria for ratings of quality and grading of strength will make judgments more transparent to those using guidelines and recommendations.

Although many grading systems to some extent meet these criteria,[1] a plethora of systems makes their use difficult for frontline clinicians. Understanding a variety of systems is neither an efficient nor a realistic use of clinicians' time. The GRADE system is used

widely: the World Health Organization, the American College of Physicians, the American Thoracic Society, UpToDate (an electronic resource widely used in North America, www.uptodate.com), and the Cochrane Collaboration are among the more than 25 organisations that have adopted GRADE. This widespread adoption of GRADE reflects GRADE's success as a methodologically rigorous, user friendly grading system.

How does the GRADE system classify quality of evidence?

To achieve transparency and simplicity, the GRADE system classifies the quality of evidence in one of four levels—high, moderate, low, and very low (box 2). Some of the organisations using the GRADE system have chosen to combine the low and very low categories. Evidence based on randomised controlled trials begins as high quality evidence, but our confidence in the evidence may be decreased for several reasons, including:

- Study limitations
- Inconsistency of results
- Indirectness of evidence
- Imprecision
- Reporting bias.

Although observational studies (for example, cohort and case-control studies) start with a "low quality" rating, grading upwards may be warranted if the magnitude of the treatment effect is very large (such as severe hip osteoarthritis and hip replacement), if there is evidence of a dose-response relation or if all plausible biases would decrease the magnitude of an apparent treatment effect.

How does the GRADE system consider strength of recommendation?

The GRADE system offers two grades of recommendations: "strong" and "weak" (though guidelines panels may prefer terms such as "conditional" or "discretionary" instead of weak). When the desirable effects of an intervention clearly outweigh the undesirable effects, or clearly do not, guideline panels offer strong recommendations. On the other hand, when the trade-offs are less certain—either because of low quality evidence or because evidence suggests that desirable and undesirable effects are closely balanced—weak recommendations become mandatory.

In addition to the quality of the evidence, several other factors affect whether recommendations are strong or weak (table 1).

Box 2 | Quality of evidence and definitions

High quality— Further research is very unlikely to change our confidence in the estimate of effect

Moderate quality— Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate

Low quality— Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate

Very low quality— Any estimate of effect is very uncertain

Factors that affect the strength of a recommendation

Factor	Examples of strong recommendations	Examples of weak recommendations
Quality of evidence	Many high quality randomised trials have shown the benefit of inhaled steroids in asthma	Only case series have examined the utility of pleurodesis in pneumothorax
Uncertainty about the balance between desirable and undesirable effects	Aspirin in myocardial infarction reduces mortality with minimal toxicity, inconvenience, and cost	Warfarin in low risk patients with atrial fibrillation results in small stroke reduction but increased bleeding risk and substantial inconvenience
Uncertainty or variability in values and preferences	Young patients with lymphoma will invariably place a higher value on the life prolonging effects of chemotherapy than on treatment toxicity	Older patients with lymphoma may not place a higher value on the life prolonging effects of chemotherapy than on treatment toxicity
Uncertainty about whether the intervention represents a wise use of resources	The low cost of aspirin as prophylaxis against stroke in patients with transient ischemic attacks	The high cost of clopidogrel and of combination dipyridamole and aspirin as prophylaxis against stroke in patients with transient ischaemic attacks

SUMMARY POINTS

Failure to consider the quality of evidence can lead to misguided recommendations; hormone replacement therapy for post-menopausal women provides an instructive example

High quality evidence that an intervention's desirable effects are clearly greater than its undesirable effects, or are clearly not, warrants a strong recommendation

Uncertainty about the trade-offs (because of low quality evidence or because the desirable and undesirable effects are closely balanced) warrants a weak recommendation

Guidelines should inform clinicians what the quality of the underlying evidence is and whether recommendations are strong or weak

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach provides a system for rating quality of evidence and strength of recommendations that is explicit, comprehensive, transparent, and pragmatic and is increasingly being adopted by organisations worldwide

Details of the GRADE working group, contributors, and competing interests appear in the version on bmj.com

1 Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches. The GRADE Working Group. *BMC Health Serv Res* 2004;4(1):38.
2 Lacchetti C, Guyatt G. Surprising results of randomized trials. In: Guyatt G, Drummond R, eds. *Users' guides to the medical literature: a manual of evidence-based clinical practice*. Chicago, IL: AMA Press, 2002.
3 American College of Physicians. Guidelines for counseling postmenopausal women about preventive hormone therapy. *Ann Intern Med* 1992;117:1038-41.
4 Humphrey LL, Chan BK, Sox HC. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. *Ann Intern Med* 2002;137:273-84.
5 Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. *JAMA* 1998;280:605-13.
6 Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA* 2002;288:321-33.
7 Echt DS, Liebson PR, Mitchell LB, Peters RW, Obias-Manno D, Barker AH, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The cardiac arrhythmia suppression trial. *N Engl J Med* 1991;324:781-8.
8 Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA* 1992;268:240-8.
9 Committee on Infectious Diseases. Aspirin and Reye syndrome. *Pediatrics* 1982;69:810-2.