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

Gordon H Guyatt, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, Holger J Schünemann and for the GRADE Working Group

BMJ 2008;336:924-926
doi:10.1136/bmj.39489.470347.AD

Updated information and services can be found at:
http://bmj.com/cgi/content/full/336/7650/924

These include:

References
This article cites 8 articles, 6 of which can be accessed free at:
http://bmj.com/cgi/content/full/336/7650/924#BIBL

Rapid responses
You can respond to this article at:
http://bmj.com/cgi/eletter-submit/336/7650/924

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top left of the article

Topic collections
Articles on similar topics can be found in the following collections

Other evidence based practice (497 articles)

Notes

To order reprints follow the "Request Permissions" link in the navigation box

To subscribe to BMJ go to:
http://resources.bmj.com/bmj/subscribers
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

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 the data came from observational studies and that because the data came from observational studies the evidence for the benefit of the drugs’ ability to reduce asymptomatic ventricular arrhythmias associated with sudden death was of low quality.7 Other drugs were licensed on the basis of observational studies of patients taking drugs for other conditions. The limitations of the evidence were not considered and the evidence was applied inappropriately.

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.
ANALYSIS

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. 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. 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, 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...
Examples of strong recommendations

Many high quality randomised trials have shown the benefit of inhaled steroids in asthma.

Examples of weak recommendations

Only case series have examined the utility of pleurodesis in pneumothorax.

Factors that affect the strength of a recommendation

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

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.

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).

**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.