ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 401–406

Journal of Clinical Epidemiology

GRADE guidelines: 3. Rating the quality of evidence

Howard Balshem[a,*], Mark Helfand[a,b], Holger J. Schünemann[c], Andrew D. Oxman[d], Regina Kunz[e], Jan Brozek[c], Gunn E. Vist[d], Yngve Falck-Ytter[f], Joerg Meerpohl[g,h], Susan Norris[i], Gordon H. Guyatt[c]

[a]*Oregon Evidence-based Practice Center, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd., Portland, OR 97239, USA*
[b]*Portland VA Medical Center, Portland, OR, USA*
[c]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[d]*Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs plass, 0130 Oslo, Norway*
[e]*Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, 4031 Basel, Switzerland*
[f]*Division of Gastroenterology, Case Medical Center and VA, Case Western Reserve University, Cleveland, OH 44106, USA*
[g]*German Cochrane Center, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, 79104 Freiburg, Germany*
[h]*Division of Pediatric Hematology and Oncology, Department of Pediatric and Adolescent Medicine, University Medical Center Freiburg, 79106 Freiburg, Germany*
[i]*Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239-3098, USA*

Accepted 30 July 2010

Abstract

This article introduces the approach of GRADE to rating quality of evidence. GRADE specifies four categories—high, moderate, low, and very low—that are applied to a body of evidence, not to individual studies. In the context of a systematic review, quality reflects our confidence that the estimates of the effect are correct. In the context of recommendations, quality reflects our confidence that the effect estimates are adequate to support a particular recommendation. Randomized trials begin as high-quality evidence, observational studies as low quality. ''Quality'' as used in GRADE means more than risk of bias and so may also be compromised by imprecision, inconsistency, indirectness of study results, and publication bias. In addition, several factors can increase our confidence in an estimate of effect. GRADE provides a systematic approach for considering and reporting each of these factors. GRADE separates the process of assessing quality of evidence from the process of making recommendations. Judgments about the strength of a recommendation depend on more than just the quality of evidence.

Keywords: Quality assessment; Body of evidence; Imprecision; Indirectness; Inconsistency; Publication bias

1. Introduction

In the two previous articles in this series, we introduced GRADE; provided an overview of the GRADE process for developing recommendations and the final outputs of that process, the evidence profile, and Summary of Findings table; and described the process for framing questions and identifying outcomes [1,2]. In this third article, we will introduce GRADE's approach to rating the quality of evidence. The goal is to provide a conceptual overview of the approach. A more detailed description, accompanied by examples, will follow in articles dealing with factors that may lead to rating down or rating up the quality of evidence [3–7].

2. What we do not mean by quality of evidence

In discussions of quality of evidence, confusion often arises between evidence and opinion and between quality of evidence and strength of recommendations. We, therefore, begin by explaining what we do not mean by quality of evidence.

3. Opinion is not evidence

In the absence of high-quality evidence, clinicians must look to lower quality evidence to guide their decisions.

The GRADE system has been developed by the GRADE Working Group. The named authors drafted and revised this article. A complete list of contributors to this series can be found on the *Journal of Clinical Epidemiology* Web site.

* Corresponding author. Oregon Evidence-based Practice Center, OHSU-BICC, 3181 SW Sam Jackson Park Rd., Portland, OR 97239, USA. Tel.: 503-220-8262 x54487; fax: 503-418-3332.

E-mail address: balshemh@ohsu.edu (H. Balshem).

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2010.07.015

Key Points

- GRADE provides a framework for assessing quality that encourages transparency and an explicit accounting of the judgments made.
- GRADE distinguishes between quality assessment conducted as part of a systematic review and that undertaken as part of guideline development.
- The optimal application of GRADE requires systematic review of the impact of alternative management strategies on all patient-important outcomes.
- Information about study limitations, imprecision, inconsistency, indirectness, and publication bias is necessary for decision makers, clinicians, and patients to understand and have confidence in the assessment of quality and estimate of effect size.

Confusion arises when, in such situations, guideline developers classify "expert opinion" as a type of evidence. Developing recommendations always requires the opinion of experts, the basis of which includes experience with patients, an understanding of biology and mechanism, and knowledge and understanding of preclinical and early clinical research as well as of the results of randomized clinical trials and observational studies. Guideline developers should always engage experts to help understand the evidence; they must also uncover and make clear the evidence that underlies the experts' opinions and rate the quality of that evidence, not the opinions that follow from the evidence and its interpretation.

An example illustrates the difference between evidence and expert opinion. Suppose that during attending rounds with medical students and residents, an endocrinologist explains the rationale for tight glycemic control in diabetes. Table 1 shows the two assertions he makes and the evidence he cites to support them. The evidence he cites for opinion 1 is exclusively his personal clinical experience. For opinion 2, he cites his own experience and refers (with no more than a general statement) to evidence from clinical research.

It seems highly plausible that opinion 1 might reasonably be based on careful observation. If patients who complain of fatigue, polyuria, or other symptoms return in a few days saying they are better, initiation of treatment is the likeliest explanation. The phenomenon of a patient who had no complaints returning, a few days later, to say how much better she is would be particularly memorable. Unfortunately, there are many other potential explanations of these observations. The endocrinologist's impression of the extent of patients' reports of benefit may be inaccurate, he may be forgetting many patients who failed to improve, or the apparent improvement in some patients may be because of natural history, placebo effects, leading questions on the part of the clinician, or the patient's desire to please. Without, at the very least, a rigorous and structured approach to data collection, we could consider the endocrinologist's report of his clinical experience (but not the opinion that he arrived at from his interpretation of that experience) as evidence from an uncontrolled case series and classify it as very low quality.

Whereas the implicit study design underlying the evidence for opinion 1 is a before−after study, opinion 2 suggests a parallel group comparison, which in this case has serious problems. If indeed his memory is accurate (patients with tighter control in his practice do achieve better outcomes), the reason may be that their success in controlling their glucose reflects differences in their underlying disease strongly associated with their likelihood of suffering complications. This risk of bias from unrecognized prognostic imbalance, as well as from the uncertainty and imprecision associated with the endocrinologist's memory of the events, would lead us again to classify his observations as very low quality evidence.

4. A particular quality of evidence does not necessarily imply a particular strength of recommendation

A second area of confusion relates to the distinction between assessing the quality of evidence and making a recommendation. Later articles in this series will provide a detailed discussion of GRADE's approach to deciding on the direction and strength of recommendations. We note here the importance of GRADE's explicit separation of the process for assessing the quality of a body of evidence from the process for making recommendations based in part on those assessments. Although higher quality evidence is more likely to be associated with strong recommendations than lower quality evidence, a particular level of quality does not imply a particular strength of recommendation. Sometimes, low or very low quality evidence can lead to a strong recommendation.

For instance, consider the decision to administer aspirin or acetaminophen to children with chicken pox. Observational studies have observed an association between aspirin administration and Reye's syndrome [8−11]. Because aspirin and acetaminophen are similar in their analgesic and antipyretic effects, the low-quality evidence regarding the potential harms of aspirin does not preclude a strong recommendation for acetaminophen.

Similarly, high-quality evidence does not necessarily imply strong recommendations. For example, faced with a first deep venous thrombosis (DVT) with no obvious provoking factor patients must, after the first months of anticoagulation, decide whether to continue taking warfarin long term. High-quality randomized controlled trials show that continuous warfarin will decrease the risk of recurrent thrombosis but at the cost of increased risk of bleeding and inconvenience [12−15]. Because patients with varying values and

Table 1
Expert opinion vs. evidence

Expert opinion	Evidence
Tight control will make a patient feel better	"In my 20 years in practice I have started treatment for newly diagnosed diabetes many times. I almost always see these patients back a week or so after starting treatment, and the great majority say they feel much better than they did before. Even a patient who denied having any complaints or symptoms will come back and say she has more energy, particularly in the afternoons, and will marvel at how much better she feels in general."
Tight control will reduce the long-term risk of developing kidney disease, neuropathy, and blindness	"I institute tight control on every patient—I believe they all deserve the best possible treatment—so I have a lot of experience with this. I have many patients who have been with me for a decade, or even several decades, and who take their medicine faithfully and have great blood sugars. These patients also have very few complications. On the other hand, I have a lot of patients who have terrible control and develop complications early on. Also, there are a lot of studies showing that tight control reduces the risk of complications."

preferences are likely to make different choices, guideline panels addressing whether patients should continue or terminate warfarin may—despite the high-quality evidence—offer a weak recommendation.

5. So what do we mean by "quality of evidence"?

GRADE distinguishes between quality assessment conducted as part of a systematic review and that undertaken in the process of guideline development. We, therefore, provide two definitions of "quality of evidence."

The optimal application of GRADE requires systematic reviews of the impact of alternative management approaches on all patient-important outcomes [1]. In the context of a systematic review, the ratings of the quality of evidence reflect the extent of our confidence that the estimates of the effect are correct. In the context of making recommendations, the quality ratings reflect the extent of our confidence that the estimates of an effect are adequate to support a particular decision or recommendation.

The reason for the different definitions is that the conduct of systematic reviews does not include processes required for making rigorous recommendations. In particular, unless the systematic review team includes members who will use the review as part of guideline development, authors of systematic reviews are, generally, not in a position to weigh the trade-offs between the desirable and undesirable consequences of adhering to a recommendation. Relevant stakeholders are in a better position to make these judgments. For example, in the DVT case described earlier, a systematic review might provide reliable estimates of the magnitude of effect and associated confidence intervals (CIs) for symptomatic thromboembolism and bleeding and the mortality associated with both of these events, but the reviewers who wrote it would not be able to provide reliable judgments about whether the benefit of warfarin treatment is worth the risk. Such judgments must also include considerations of values, cost, and pertinent stakeholder input.

On the other hand, a guideline (or a clinician applying the evidence from a systematic review) must assess the quality of the evidence in the context of the decision regarding anticoagulation. In considering this trade-off, a guideline panel must decide whether or not to recommend anticoagulation (and the strength of that recommendation) in light of the effect on the risk of symptomatic thromboembolism, their confidence in the effect estimates, and the corresponding risks and confidence in estimates of serious bleeding. Although the processes for assessing quality are the same, authors of systematic reviews and authors of guidelines will apply the criteria differently. We will highlight this different application of criteria in the fifth article in this series, which addresses the assessment of precision in rating the quality of the evidence [5].

6. Quality in GRADE means more than risk of bias

In the clinical epidemiological literature, when used at all, "quality" commonly refers to a judgment on the internal validity (i.e., risk of bias) of an individual study. To arrive at a rating, reviewers consider features in controlled trials such as randomization, allocation concealment, blinding, and use of intention to treat analysis. In observational studies, they consider appropriate measurement of exposure and outcome as well as appropriate control of confounding; in both controlled trials and observational studies, they consider loss to follow-up and may consider other aspects of design, conduct, and analysis that influence the risk of bias.

GRADE judgments refer not to individual studies but to a body of evidence, and quality, as used in GRADE, means more than risk of bias. A body of evidence (for instance, a number of well-designed and executed trials) may be associated with a low risk of bias, but our confidence in effect estimates may be compromised by a number of other factors (imprecision, inconsistency, indirectness, and publication bias). There are also factors, particularly relevant to observational studies, that may lead to rating up quality, including the magnitude of treatment effect and the presence of a dose−response gradient.

GRADE's specific uses of the terms "quality" and "risk of bias" (labeled "study limitations" in previous GRADE publications) require authors to take care in using these terms when they describe their findings and reasoning in

Table 2
Significance of the four levels of evidence

Quality level	Current definition	Previous definition
High	We are very confident that the true effect lies close to that of the estimate of the effect	Further research is very unlikely to change our confidence in the estimate of effect
Moderate	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate
Low	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect	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	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect	Any estimate of effect is very uncertain

the context of a systematic review or guideline. Well-conducted studies may be part of a body of evidence rated low quality because they only provide indirect or imprecise evidence for the question of interest. Although clinical epidemiologists and others have attributed other meanings to the word "quality" (typically risk of bias), we believe the meaning described here corresponds more closely to the common and nontechnical understanding of "quality."

7. GRADE specifies four categories for the quality of a body of evidence

Although the quality of evidence represents a continuum, the GRADE approach results in an assessment of the quality of a body of evidence as high, moderate, low, or very low. Table 2 presents what GRADE means by each of these four categories and contrasts their current definition with the previous definition [16], which focused on the implications of the levels of evidence for future research (the lower the quality, the more likely further research would change our confidence in the estimates, and the estimates themselves). The earlier characterization has been criticized—we believe legitimately—because there are many situations in which we cannot expect higher quality evidence to be forthcoming. We, nevertheless, consider the prior characterization of quality to provide an alternative under circumstances when obtaining new compelling evidence is plausible.

8. Arriving at a quality rating

When we speak of evaluating quality, we are referring to an overall rating for each important outcome across studies. As discussed in the previous article in this series that addressed the framing of the question [2], before assessing the quality of the evidence, systematic reviewers and guideline developers should identify all potential patient-important outcomes, including benefits, harms, and costs. Reviewers will then assess the quality of evidence for each important outcome.

Table 3 summarizes GRADE's approach to rating the quality of evidence, which begins with the study design (trials or observational studies) and then addresses five reasons to possibly rate down the quality of evidence and three to possibly rate up the quality. Subsequent articles in this series will address, in detail, the meaning and use of each of these criteria. Here, we discuss why these criteria, in particular, have been identified as important in assessing the quality of a body of evidence.

Table 3
A summary of GRADE's approach to rating quality of evidence

Study design	Initial quality of a body of evidence	Lower if	Higher if	Quality of a body of evidence
Randomized trials	High ⇒	Risk of Bias −1 Serious −2 Very serious Inconsistency −1 Serious −2 Very serious Indirectness −1 Serious −2 Very serious Imprecision −1 Serious −2 Very serious Publication bias −1 Likely −2 Very likely	Large effect +1 Large +2 Very large Dose response +1 Evidence of a gradient All plausible residual confounding +1 Would reduce a demonstrated effect +1 Would suggest a spurious effect if no effect was observed	High (four plus: ⊕⊕⊕⊕) Moderate (three plus: ⊕⊕⊕◯) Low (two plus: ⊕⊕◯◯) Very low (one plus: ⊕◯◯◯)
Observational studies	Low ⇒			

9. Rationale for using GRADE's definition of quality

To be useful to decision makers, clinicians, and patients, systematic reviews must provide not only an estimate of effect for each outcome but also the information needed to judge whether these estimates are likely to be correct. What information about the studies in a review affects our confidence that the estimate of an effect is correct?

To answer this question, consider an example. Suppose you are told that a recent Cochrane review reported that, in patients with chronic pain, the number needed to treat (NNT) for clinical success with topical salicylates was 6 (95% CI = 4−13) compared with placebo. What additional information would you seek to help you decide whether to believe this estimate and how to apply it?

The most obvious questions might be the following: how many studies were pooled to get this estimate; how many patients did they include; and how wide were the CIs around the effect estimate? Were they randomized controlled trials? Did the studies have important limitations, such as lack of blinding or large or differential loss to follow-up in the compared groups? The questions thus far relate to GRADE categories of imprecision and risk of bias.

But there are also other important questions. Is there evidence that more studies of this treatment were conducted, but some were inaccessible to the reviewers? If so, how likely is it that the results of the review reflect the overall experience with this treatment? Did the trials have similar or widely varying results? Was the outcome measured at an appropriate time, or were the studies too short in duration to have much relevance? What part of the body was involved in the interventions (and thus, to what part of the body can we confidently apply these results)? These latter questions refer to the GRADE categories of publication bias, inconsistency, and indirectness. Without answers to (or at least information about) these questions, it is not possible to determine how much confidence to attach to the reported NNT and CIs.

GRADE identified its five categories—risk of bias, imprecision, inconsistency, indirectness, and publication bias—because they address nearly all issues that bear on the quality of evidence. For any given question, moreover, information about each of these categories is likely to be essential to judge whether the estimate is likely to be correct. These categories were arrived at through a case-based process by members of GRADE, who identified a broad range of issues and factors related to the assessment of the quality of studies. All potential factors were considered, and through an iterative process of discussion and review, concerns were scrutinized and solutions narrowed by consensus to these five categories.

GRADE's approach to quality implies that every systematic review should provide information about each of the categories (and any other pertinent issues in a particular case). Decision makers, whether they are guideline developers or clinicians, find it difficult to use a systematic review that does not provide this information. Good systematic reviews and clinical practice guidelines have commonly emphasized appraisal of the risk of bias (study limitations) using explicit criteria. Often, however, the focus has been on assessments across outcomes for each study rather than on each important outcome across studies. Assessment of other factors that determine how much confidence can be placed in estimates of effect has often been lacking. Before the adoption of GRADE, standards for reporting systematic reviews have not made clear how this information should be presented. GRADE provides a structure for systematic reviews and clinical practice guidelines to ensure they address the key questions that are pertinent to rating the quality of the evidence for all outcomes relevant to a particular question in a consistent systematic manner.

10. Conclusion

In closing, we caution against a mechanistic approach toward the application of the criteria for rating the quality of the evidence up or down. Although GRADE suggests the initial separate consideration of five categories of reasons for rating down the quality of evidence, and three categories for rating up, with a yes/no decision regarding rating up or down in each case, the final rating of overall evidence quality occurs in a continuum of confidence in the validity, precision, consistency, and applicability of the estimates. Fundamentally, the assessment of evidence quality is a subjective process, and GRADE should not be seen as obviating the need for or minimizing the importance of judgment or as suggesting that quality can be objectively determined.

As we repeatedly stress throughout this series, use of GRADE will not guarantee consistency in assessment, whether of the quality of evidence or of the strength of recommendations. There will be cases in which competent reviewers will have honest and legitimate disagreement about the interpretation of evidence. In such cases, the merit of GRADE is that it provides a framework that guides one through the critical components of this assessment and an approach to analysis and communication that encourages transparency and an explicit accounting of the judgments involved.

References

[1] Guyatt GH, Oxman AD, Kunz R, Vist GE, Brozek J, Norris S, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383−94 [in this issue].

[2] Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist GE, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64:395−400 [in this issue].

[3] Guyatt GH, Oxman AD, Vist GE, Kunz R, Brozek J, Alonso-Coello, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol 2011;64:407−15.

[4] Guyatt GH, Oxman AD, Montori V, Vist GE, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence - publication bias. J Clin Epidemiol. In press.
[5] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines: 6. Rating the quality of evidence—imprecision (random error). J Clin Epidemiol. In press.
[6] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. J Clin Epidemiol. In press.
[7] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. J Clin Epidemiol. In press.
[8] Waldman RJ, Hall WN, McGee H, Van Amburg G. Aspirin as a risk factor in Reye's syndrome. JAMA 1982;247:3089−94.
[9] Starko KM, Ray CG, Dominguez LB, Stromberg WL, Woodall DF. Reye's syndrome and salicylate use. Pediatrics 1980;66:859−64.
[10] Halpin TJ, Holtzhauer FJ, Campbell RJ, Hall LJ, Correa-Villasenor A, Lanese R, et al. Reye's syndrome and medication use. JAMA 1982;248: 687−91.
[11] Hurwitz ES, Barrett MJ, Bregman D, Gunn WJ, Pinsky P, Schonberger LB, et al. Public health service study of Reye's syndrome and medications: report of the main study. JAMA 1987;257: 1905−11.
[12] Kearon C, Gent M, Hirsh J, Weitz J, Kovacs MJ, Anderson DR, et al. A comparison of three months of anticoagulation with extended anticoagulation for a first episode of idiopathic venous thromboembolism. N Engl J Med 1999;340:901.
[13] Campbell IA, Bentley DP, Prescott RJ, Routledge PA, Shetty HGM, Williamson IJ. Anticoagulation for three versus six months in patients with deep vein thrombosis or pulmonary embolism, or both: randomised trial. BMJ 2007;334:674.
[14] Kearon C, Ginsberg JS, Anderson DR, Kovacs MJ, Wells P, Julian JA, et al. Comparison of 1 month with 3 months of anticoagulation for a first episode of venous thromboembolism associated with a transient risk factor. J Thromb Haemost 2004;2:743−9.
[15] Agnelli G, Prandoni P, Santamaria MG, Bagatella P, Iorio A, Bazzan M, et al. Three months versus one year of oral anticoagulant therapy for idiopathic deep venous thrombosis. N Engl J Med 2001; 345:165.
[16] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008; 336:924−6.