METHODOLOGY

Understanding GRADE: an introduction
Gabrielle Goldet and Jeremy Howick

Department of Primary Health Sciences, University of Oxford, Oxford, UK

Keywords
Evidence-based medicine; GRADE; randomized controlled trial; systematic review.

Correspondence
Jeremy Howick, Department of Primary Care Health Sciences, New Radcliffe House, University of Oxford, Oxford OX2 2GG, UK. Tel: 01865 289 258 Fax: 01865 289 287 Email: jeremy.howick@phc.ox.ac.uk

Abstract

Objective: Grading of recommendations, assessment, development, and evaluations (GRADE) is arguably the most widely used method for appraising studies to be included in systematic reviews and guidelines. In order to use the GRADE system or know how to interpret it when reading reviews, reading several articles and attending a workshop are required. Moreover, the GRADE system is not covered in standard medical textbooks. Here, we explain GRADE concisely with the use of examples so that students and other researchers can understand it.

Background: In order to use or interpret the GRADE system, reading several articles and attending a workshop is currently required. Moreover, the GRADE system is not covered in standard medical textbooks.

Methods: We read, synthesized, and digested the GRADE publications and contacted GRADE contributors for explanations where required. We composed a digested version of the system in a concise way a general medical audience could understand.

Results: We were able to explain the GRADE basics clearly and completely in under 1500 words.

Conclusions: While advanced critical appraisal requires judgment, training, and practice, it is possible for a non-specialist to grasp GRADE basics very quickly.

Introduction

The Grading of recommendations, assessment, development, and evaluations (GRADE) system is emerging as the dominant method for appraising controlled studies and making recommendations for systematic reviews and guidelines (1–12). Reading the series of publications dedicated to explaining GRADE (2–12) and perhaps attending a workshop are usually required to grasp the GRADE system in sufficient detail to be able to use it for critical appraisal. There are no concise explanations of GRADE intended for a general medical audience and medical textbooks do not include explanations of GRADE. GRADE may thus elude students, doctors, and policy makers who do not have the time to study the literature or attend the workshops. Yet anyone reading a systematic review or evaluating guideline recommendations needs to understand GRADE in order to appraise the review or guideline. Taking all appraisal skills out of the hands of clinicians and into the hands of expert reviewers is undesirable from an Evidence-Based Medicine perspective. Therefore a clear, brief, and accurate description of GRADE is required.

What is GRADE?

GRADE is a method used by systematic reviewers and guideline developers to assess the quality of evidence and decide
whether to recommend and intervention (1). GRADE differs from other appraisal tools for three reasons: (i) because it separates quality of evidence and strength of recommendation, (ii) the quality of evidence is assessed for each outcome, and (iii) observational studies can be ‘upgraded’ if they meet certain criteria.

**Using GRADE**

The GRADE method involves five distinct steps that we explain here with examples.

**Step 1**

Assign an a priori ranking of ‘high’ to randomized controlled trials and ‘low’ to observational Studies. Randomized controlled trials are initially assigned a higher grade because they are usually less prone to bias than observational studies (13, 14).

**Step 2**

‘Downgrade’ or ‘upgrade’ initial ranking. It is common for randomized controlled trials and observational studies to be downgraded because they suffer from identifiable bias (15). Also, observational studies can be upgraded when multiple high-quality studies show consistent results (16–19).

**Reasons to ‘downgrade’**

- **Risk of bias**
  - Lack of clearly randomized allocation sequence
  - Lack of blinding
  - Lack of allocation concealment
  - Failure to adhere to intention-to-treat analysis
  - Trial is cut short
  - Large losses to follow-up.

There is strong evidence supporting the view that lack of randomization, lack of allocation concealment, and lack of blinding biases results (20). Intention-to-treat analysis is also important to avoid bias arising when those dropping out of trials due to harmful effects are not accounted for (21). Similarly, interim results from trials that are cut short often lead to overestimated effect sizes (22). Large losses due to follow-up also lead to exaggerated effect estimates. For example, in a randomized controlled trial comparing weight loss programs in obese participants, 27% of participants were lost to follow-up in a year. There is a putative positive bias conferred on this study as participants benefiting from the treatment were more likely to stay in the trial (23).

- **Inconsistency** when there is significant and unexplained variability in results from different trials and could not assert whether treatment benefits outweighed the harms (24).

- **Indirectness of evidence** can refer to several things.
  - An indirect comparison of two drugs. If there are no trials comparing drugs A and B directly, we infer a comparison based on separate trials comparing drug A with placebo and drug B with placebo. However the validity of such an inference depends on the often unjustified assumption that the two trial populations were similar.
  - An indirect comparison of population, outcome, or intervention, for example, the studies in the review investigated one intervention, in one population with a certain outcome, but the conclusions of the studies are intended to be applicable for related interventions, populations, or outcomes. For example, the American College of Chest Physicians (ACCP) downgraded the quality of the evidence for compression stocking use in trauma patients from high to moderate because all the randomized controlled trials had been performed on the general population (25).

- **Imprecision** when wide confidence intervals mar the quality of the data.

- **Publication bias** when studies with ‘negative’ findings remain unpublished. This can bias the review outcome (26). For example, Turner et al examined all antidepressant trials registered by the US Food and Drug Administration. Of 38 trials with positive results, all but one was published, whereas of the 36 trials with negative results, 22 were unpublished (27)—a systematic review of all published studies would give a biased result.

**Reasons to ‘upgrade’**

- **Large effect** when the effect is so large that bias common to observational studies cannot possibly account for the result. For example, observational studies examining the effect of antithrombotic therapy on thromboembolic events in patients having undergone cardiac valve replacement revealed an 80% relative risk reduction, which was used by the ACCP to rank the quality of evidence as high (28).

- **Dose–response relationship** when the result is proportional to the degree of exposure. For example, the Nurses’ Health Study demonstrated a strong negative correlation between physical activity and stroke risk—that is the more physically active the nurses were, the smaller the risk of stroke (29).

- **All plausible biases only reducing an apparent treatment effect** when all possible confounders would only diminish the observed effect. It is likely thus that the
actual effect is larger than the data suggests. For example, a systematic review of studies in 26,000 hospitals found mortality was higher in for-profit private hospitals compared to non-profit public hospitals. Investigators suspected teaching hospitals might have confounded the study because they are nonprofit and were suspected to have low mortality. When teaching hospitals were removed from the analysis, however, the relative mortality of for-profit hospitals actually increased (30).

**Step 3**
Assign final grade for the quality of evidence as ‘high’, ‘moderate’, ‘low’, or ‘very low’ for all the critically important outcomes (Box 1) (31). Upgrading and downgrading requires careful judgment and thought. A rule of thumb is to move up or down one category per issue. For example, the quality of evidence for an outcome studied in randomized controlled trials that initially starts as ‘high’ might move to ‘moderate’ because of high risk of bias in the included studies, and further down to ‘low’ if there was significant unexplained heterogeneity between the trials, and even further down to ‘very low’ if there were few events leading to confidence intervals including both important benefits and harms.

**Box 1 Final GRADE ranking**
- **High** we are very confident that the effect in the study reflects the actual effect.
- **Moderate** we are quite confident that the effect in the study is close to the true effect, but it is also possible it is substantially different.
- **Low** the true effect may differ significantly from the estimate.
- **Very low** the true effect is likely to be substantially different from the estimated effect.

---

**Step 4**
Consider other factors that impact on the strength of recommendation for a course of action. High-quality evidence does not always imply a strong recommendation. Recommendations must consider factors besides to the quality of evidence (3). The first factor to consider is the balance between desirable and undesirable effects. Some interventions, such as antibiotics to prevent urinary infections associated with certain urologic procedures, have clear benefits and few side effects and therefore a strong recommendation is uncontroversial (32, 33). In cases where the benefit to harm ratio is less clear, then patient values and preferences, as well as costs, need to be considered carefully (Figure 1).

**Step 5**
Make a ‘strong’ or ‘weak’ recommendation (34). When the net benefit of the treatment is clear, patient values and circumstances are unlikely to affect the decision to take a treatment and a strong recommendation is appropriate. For example, high-dose chemotherapy for testicular cancer is often highly recommended for the increased likelihood of survival in spite of chemotherapy’s toxicity (32, 35). Otherwise, if the evidence is weak or the balance of positive and negative effects is vague, a weak recommendation is appropriate. Consider the following example (borrowed from the GRADE literature) concerning whether to recommend screening for melanoma in primary care in the United States (36, 37). The baseline risk for this population is 13.3 per 100,000, and there is weak evidence for the accuracy of screening and the outcome of lethal melanoma; potential harms from screening is lacking yet likely to include harmful consequences of false positive tests (38). Based on this lack of strong evidence, the recommendation to screen or not to screen will not be strong, and may require input from patients, friends, and family members about the

---

**Figure 1** How GRADE is used to make recommendations; steps 1 to 3 are repeated for each critical outcome.
different treatment options. Some people might choose to recommend not screening because of the lack of strong evidence for benefit and potential harms. Others might recommend screening because they value the potential benefits more highly (34).

Summary

GRADE is becoming increasingly popular for guideline developers and systematic reviewers. It is thus important for anyone (students, clinicians, and policy makers) to understand GRADE. Advanced critical appraisal (using GRADE or any other system) requires sound judgment, training, and practice. However it is possible to grasp GRADE basics using this concise and accurate summary.

Acknowledgement

We are very grateful to Gunn Vist, who provided expert advice on the GRADE system and editorial suggestions.

References


