Grading the quality of evidence and the strength of recommendations: The GRADE approach

Bassam Elhelw, MRCOG, MFFP

Middle East Fertility Center, Dokki, Giza, Egypt.

Clinical recommendations that are based on medical evidence can be of different quality. Sources of evidence range from small laboratory studies or case reports to well-designed large clinical studies that have minimized bias to a great extent. Since poor quality evidence can lead to recommendations that are not in patients’ best interests, it is essential to know whether a recommendation is strong or weak.

Grading evidence and recommendations is not a new issue. Since the 1970s a growing number of organizations have employed various systems to grade the quality (level) of evidence and the strength of recommendations (1-9). Unfortunately, different organizations use different systems to grade the quality of evidence and the strength of recommendations. The same evidence and recommendation could be graded as II-2, B; C+, 1; or strong evidence, strongly recommended depending on which system is used. This is confusing and impedes effective communication (10).

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group began in the year 2000 as an informal collaboration of people with an interest in tackling the shortcomings of present grading systems in health care. The mission of the GRADE working group is to help resolve the confusion among the different systems of rating evidence and recommendations (11). The group has wide representation from many organizations including the Agency for Healthcare Research and Quality in the US, the National Institute for Clinical Excellence for England and Wales, and the World Health Organization. Developing a new uniform rating system is challenging because all systems have limitations and because many organizations have invested a great deal of time and effort to develop their rating systems and is understandably reluctant to adopt a new system (11, 12).

Comparison of GRADE system and other grading systems

Definitions: GRADE uses explicit definitions e.g. the quality of evidence indicates the extent to which one can be confident that an estimate of effect is correct, the strength of a recommendation indicates the extent to which one can be confident that adherence to the recommendation will do more good than harm. These explicit definitions make clear what grades indicate and what should be considered in making these judgments. Other systems use implicit definitions of quality (level) of evidence and strength of recommendation.

Judgments: Sequential, explicit judgments regarding which outcomes are important, quality of evidence for each important outcome, overall quality of evidence, balance between benefits and harms, and value of incremental benefits; clarifies each of these judgments and reduces risks of
introducing errors or bias that can arise when they are made implicitly

**Key components of quality of evidence:** These are not considered for each important outcome, and judgments about quality of evidence are often based on study design alone in other grading systems. GRADE uses systematic and explicit consideration of study design, study quality, consistency, and directness of evidence in judgments about quality of evidence.

**Other factors that can affect quality of evidence:** GRADE ensures explicit consideration of imprecise or sparse data, reporting bias, strength of association, evidence of a dose-response gradient, and plausible confounding. These factors were not explicitly taken into account by other systems.

**Overall quality of evidence:** in other systems overall quality of evidence was implicitly based on the quality of evidence for benefits, in GRADE it is based on the lowest quality of evidence for any of the outcomes that are critical to making a decision, leading to a reduced likelihood of mislabeling overall quality of evidence when evidence for a critical outcome is lacking.

**Relative importance of outcomes:** In order to ensure appropriate consideration of each outcome when grading overall quality of evidence and strength of recommendations; GRADE advocates explicit judgments about which outcomes are critical, which ones are important but not critical, and which ones are unimportant and can be ignored. Other systems consider relative importance of outcomes implicitly.

**Balance between health benefits and harms:** Explicit consideration of trade-offs between important benefits and harms, the quality of evidence for these, translation of evidence into specific circumstances, and certainty of baseline risks clarifies and improves transparency of judgments on harms and benefits. Balance between benefits and harm is not explicitly considered in other systems.

**Whether incremental health benefits are worth the costs:** In GRADE explicit consideration after first considering whether there are net health benefits ensures that judgments about value of net health benefits are transparent.

**Summaries of evidence and findings:** These are inconsistently presented in other systems. Consistent GRADE evidence profiles, including quality assessment and summary of findings ensures that all panel members base their judgments on same information and that this information is available to others.

**Extent of use:** Other grading systems are seldom used by more than one organization and little, if any empirical evaluation. In GRADE system there is international collaboration across wide range of organizations in development and evaluation to achieve a system that is more sensible, reliable, and widely applicable.

It is worth mentioning that most other grading systems do not include any of these advantages, although some may incorporate some of these advantages.

**The Quality of Evidence**

The GRADE system classifies quality of evidence into 1 of 4 levels: high, moderate, low, and very low.

1. **High:** further research is very unlikely to change our confidence in the estimate of effect.
2. **Moderate:** further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
3. **Low:** 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.
4. **Very low:** any estimate of effect is very uncertain.

Evidence based on randomized controlled trials (RCTs) begins with a top rating; GRADE takes into account, however, that not all RCTs are alike, and that many factors may decrease the quality of evidence e. g. Quality decreases if most of the evidence comes from RCTs with serious
methodological flaws such as lack of allocation concealment or blinding, or large loss to follow up. Inconsistency of results downgrades the quality of evidence; our confidence in the estimates of benefit or risk is weaker if some studies show substantial effects and other apparently similar studies show no effect at all. Indirectness of evidence may compromise the quality of evidence. Evidence is indirect if there are no head to head comparisons between therapeutic alternatives. Sparse evidence; when total sample size is small, and outcome events are few, our uncertainty about estimates of benefit and risk increase, and reporting bias (including publication bias).

On the other hand, observational studies (e.g., cohort studies) start with a ‘‘low quality’’ rating, but they may be graded upwards if the magnitude of the treatment effect is very large (e.g., hip replacement for severe hip osteoarthritis), if there is evidence of a dose response relation, or if all apparent confounders would decrease the magnitude of the treatment effect.

The Strength of Recommendations

The GRADE system offers two levels of recommendations that are easy to put into practice:

1. Strong: When the benefits of an intervention clearly outweigh its risks and burden, or clearly do not, strong recommendations are warranted.
2. Weak: When the tradeoff between benefits and risks is less certain, either because of low quality evidence or because high quality evidence suggests benefits and risks are closely balanced, weak recommendations become appropriate.

Although recommendations may still differ according to local circumstances and differing values and preferences, the evidentiary basis, and the quality of evidence rating (using the GRADE approach) will be uniform and guidelines will be framed using the simple, clinically applicable GRADE system of strong or weak recommendations. The GRADE working group has developed a free software application (GRADEpro) that facilitates the use of the approach and allows the development of summary tables. The software will soon be released to the public at http://www.gradeworkinggroup.org. This development has enormous implications for efficiency, improved communication, and optimal clinical decision making.

REFERENCES