An emerging consensus on grading recommendations?

Clinical practice guidelines have improved in quality over the past 10 years by adhering to a few basic principles, such as conducting thorough systematic reviews of relevant evidence and grading the recommendations and the quality of the underlying evidence. The large number of systems of measuring the quality of evidence and recommendations that have emerged are, however, confusing (1).

The mission of the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) working group is to help resolve the confusion among the different systems of rating evidence and recommendations. The group has wide representation from many organizations, including the Agency for Healthcare Research and Quality in the United States, the National Institute for Clinical Excellence for England and Wales, and the World Health Organization. Developing a new uniform ratings system is challenging because all systems have limitations and because many organizations have invested a great deal of time and effort to develop their ratings systems and are understandably reluctant to adopt a new system.

The GRADE working group first published the results of its work in 2004 in *BMJ* (2). A simpler, clinically oriented description will soon be published (3). GRADE has taken care to ensure its suggested system is simple to use and applicable to a wide variety of clinical recommendations that span the full spectrum of medical specialties and clinical care.

The GRADE system classifies recommendations in 1 of 2 levels—strong and weak—and quality of evidence into 1 of 4 levels—high, moderate, low, and very low. Evidence based on randomized controlled trials (RCTs) begins with a top rating on GRADE's 4-level quality-of-evidence classification (Table 1). GRADE takes into account, however, that not all RCTs are alike and that limitations of individual RCTs may compromise the quality of their evidence (Table 2).

First, 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. A second reason for downgrading is inconsistency of results—our confidence in estimates of benefit or risk is weaker if some studies show substantial effects and other apparently similar studies show no effect at all.

Indirectness may compromise the quality of evidence. Evidence is indirect if there are no head-to-head comparisons between therapeutic alternatives. For instance, drug benefit plans or formularies have to choose between funding of a number of bisphosphonates, including alendronate and risedronate, for prevention of osteoporotic fractures. Unfortunately, the decision must be made by comparing trials that evaluate alendronate against placebo, and risedronate against placebo, rather than directly comparing alendronate and risedronate. Evidence may also be indirect if differences exist in population (we are interested in valvular atrial fibrillation, but all RCTs are of nonvalvular atrial fibrillation), intervention (we'd like to know about relatively low-dose angiotensin-converting enzyme inhibition, but all trials are of higher dose), or outcome (we'd like to know about long-term effectiveness, but all trials have only short follow-up).

When total sample size is small and outcome events are few, our uncertainty about estimates of benefit and risk increases. GRADE continues to debate the appropriate thresholds for decreasing strength of inference: When are confidence intervals too wide? How many events are too few?

Table 1. Quality of evidence and definitions

Grade	Definition
High	Further research is very unlikely to change our confidence in the estimate of effect
Moderate	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate
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
Very low	Any estimate of effect is very uncertain

While observational studies (e.g., cohort studies) start with a "low" quality rating, they may be graded upward 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 (Table 2). For example, a systematic review showed higher mortality in for-profit than in not-for-profit hospitals (4). This result was found despite the fact that for-profit hospitals usually admit healthier patients with a higher socioeconomic status and have more resources at their disposal. These potential confounders, if anything, favor for-profit hospitals. If such confounders were taken into account, the magnitude of effect favoring not-for-profit hospitals would be even larger.

As noted, the GRADE system offers 2 levels of recommendations: strong and weak. When the benefits of an intervention clearly outweigh its risks and burden, or clearly do not, strong recommendations are warranted. On the other hand, when the tradeoff between benefits and risks is less certain, either because of low-quality evidence or because high-quality evidence suggests that benefits and risks are closely balanced, weak recommendations become appropriate.

This 2-level approach is easy to put into practice. For strong recommendations, in which it is clear that benefits far outweigh risks or that risks far outweigh benefits, virtually all patients will make the same choice (e.g., aspirin in the setting of acute myocardial infarction). In such instances, physicians can confidently

Table 2. Factors in deciding on confidence in estimates of benefits, risks, burden, and costs

Factors that may decrease the quality of evidence	Poor quality of planning and implementation of the available RCTs suggesting high likelihood of bias Inconsistency of results Indirectness of evidence Sparse evidence Reporting bias (including publication bias)
Factors that may increase the quality of evidence based on observational studies	Large magnitude of effect All plausible confounding would reduce a demonstrated effect Dose-response gradient

recommend treatment. For weak recommendations, different patients may choose different approaches to treatment. One example is the use of hormone replacement therapy for menopausal hot flashes. Under these circumstances, clinicians must know the evidence and communicate the evidence to their patients, or conduct a detailed inquiry to ensure their recommendations are consistent with patients' values and preferences (5). Beyond the quality of the evidence, a number of other factors may bear on whether recommendations are strong or weak (Table 3).

The GRADE system is rigorous in its methodology, yet practical to use. It is neither too complex nor misleadingly simple. The Cochrane Collaboration is moving to adopt the GRADE approach for the rating of methodological quality, and the revised Quality of Reporting of Systematic Reviews (QUOROM) statement is likely to endorse the approach. The Endocrine Society is the first North American organization to adopt GRADE for its recommendations, while another important organization, the American College of Chest Physicians (ACCP), has adopted a slightly modified version of GRADE. Other organizations, such as the American Thoracic Society and the BMJ Publishing Group, which publishes *Clinical Evidence*, will be exploring possible use of the GRADE approach.

The ACCP modification, which collapses the low- and very-low-quality categories, represents a simplification that may be attractive to groups providing recommendations primarily for clinical practice (rather than, for instance, public health interventions). In a particularly significant development, the popular electronic medical text UpToDate is moving to formal structured recommendations using this particular modification of the GRADE approach.

The leading American and European urology associations have, for some years, been among the leaders in evidence-based guidelines. But, as in every other area, individual urology organizations have collected and summarized evidence separately and come up with separate approaches to developing and grading recommendations. This is not only unnecessarily time-consuming but also confusing for consumers of guidelines.

American, European, and Asian urology organizations all plan to adopt GRADE for their recommendations. This alliance is also likely to result in pooled resources, use of standard approaches to collecting and summarizing evidence relevant to urology practice, and sharing these evidence summaries across groups. Recommendations may still differ according to local circumstances and differing values and preferences, but the evidentiary basis and the quality of evidence rating (using the GRADE approach) will be uniform. Guidelines will be framed using the simple, clinically applicable GRADE system of strong or weak recommendations.

This development has enormous implications for efficiency, improved communication, and optimal clinical decision-making. If the urology community can overcome political barriers and achieve this eminently sensible revolution in knowledge management and expert guidance, why not orthopedic surgery, respirology, nephrology, and even cardiology?

Most of the developments we have described are still evolving. Perhaps our vision of their eventual outcome is overly sanguine. If we can maintain momentum, however, GRADE will do more than achieve the worthy and important goal of standardizing systems of grading quality of evidence and recommendations for clinical practice. GRADE may facilitate the evolution toward a world in which expert recommendations for front-line clinicians uniformly adhere to principles of evidence management and guideline development that flow from the intellectual movement we call evidence-based medicine.

Table 3. Factors in deciding on a strong or weak recommendation*

Issue	Example
Methodological quality of the evidence supporting estimates of likely benefit and likely risk, inconvenience, and costs	Many high-quality RCTs have shown the benefit of inhaled steroids in asthma, while only case series have examined the utility of pleurodesis in pneumothorax
Importance of the outcome that treatment prevents	Preventing postphlebitic syndrome with thrombolytic therapy in DVT in contrast to preventing death from PE
Magnitude of treatment effect	Clopidogrel vs aspirin leads to a smaller stroke reduction in TIA (RRR 8.7% [6]) than anticoagulation vs placebo in atrial fibrillation (AF) (RRR 68%)
Precision of estimate of treatment effect	ASA vs placebo in AF has a wider confidence interval than ASA for stroke prevention in patients with TIA
Risks associated with therapy	ASA and clopidogrel for anticoagulation therapy in acute coronary syndromes have a higher risk for bleeding than ASA alone
Burdens of therapy	Taking adjusted-dose warfarin is associated with a higher burden than taking aspirin; warfarin requires monitoring the intensity of anticoagulation and a relatively constant dietary vitamin K intake
Risk for target event	Some surgical patients are at very low risk for postoperative DVT and PE, while other surgical patients have considerably higher rates of DVT and PE
Costs	Clopidogrel has much higher cost in patients with TIA than aspirin
Varying values	Most young, healthy persons will put a high value on prolonging their lives (and thus incur suffering to do so); older and infirm persons are likely to vary in the value they place on prolonging their lives (and may vary in the suffering they are willing to incur)

*AF = atrial fibrillation; ASA = acetylsalicylic acid; DVT = deep venous thrombosis; PE = pulmonary embolism; RCT = randomized controlled trial; RRR = relative risk reduction; TIA = transient ischemic attack.

Gordon Guyatt, MSc, MD
McMaster University
Hamilton, Ontario, Canada

Gunn Vist, PhD
Norwegian Knowledge Centre for the Health Services
Oslo, Norway

Yngve Falck-Ytter, MD
Case Western Reserve University
Cleveland, Ohio, USA

Regina Kunz, MD, MSc, PhD
Institute for Clinical Epidemiology
Basel, Switzerland

Nicola Magrini, MD
Centre for Evaluation of the Effectiveness of Health Care
Modena, Italy

Holger Schunemann, MD, PhD
McMaster University and Italian National Cancer Institute Regina Elena
Hamilton, Ontario, Canada and Rome, Italy

References

1. **Schunemann HJ, Best D, Vist G, Oxman AD.** CMAJ. 2003;169:677-80.
2. **Atkins D, Best D, Briss PA, et al.** BMJ. 2004;328:1490.
3. **Guyatt GG, Bauman M, Adrizzo-Harris D, et al.** Chest. 2005. In press.
4. **Devereaux PJ, Choi PT, Lacchetti C, et al.** CMAJ. 2002;166:1399-406.
5. **Charles C, Whelan T, Gafni A.** BMJ. 1999;319:780-2.
6. **CAPRIE Steering Committee.** Lancet. 1996;348:1329-39.