RATING QUALITY OF EVIDENCE AND STRENGTH OF RECOMMENDATIONS

GRADE: what is "quality of evidence" and why is it important to clinicians?

Guideline developers use a bewildering variety of systems to rate the quality of the evidence underlying their recommendations. Some are facile, some confused, and others sophisticated but complex

In 2004 the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group presented its initial proposal for patient management.[1] In this second of a series of five articles focusing on the GRADE approach to developing and presenting recommendations we show how GRADE has built on previous systems to create a highly structured, transparent, and informative system for rating quality of evidence.

A guideline's formulation should include a clear question

Any question addressing clinical management has four components: patients, an intervention, a comparison, and the outcomes of interest.[2] For example, consider the following: in patients with pancreatic carcinoma undergoing surgery, what is the impact of a modified resection that preserves the pylorus compared with a standard wide tumour resection—variations of the Whipple procedure—on short term and long term mortality, blood transfusions, bile leaks, hospital stay, and problems with gastric emptying?

Guideline developers should address the importance of their outcomes

GRADE challenges guideline developers to specify all outcomes that are of importance to patients as they begin the guideline development process, and to differentiate the critical outcomes from the important but not critical ones.[3] Figure 1 presents a hierarchy of patient important outcomes regarding the impact of phosphate lowering drugs in patients with renal failure. GRADE suggests a nine point scale to judge importance. The upper end of the scale, 7 to 9, identifies outcomes of critical importance for decision making. Ratings of 4 to 6 represent outcomes that are important but not critical. Ratings of 1 to 3 are items of limited importance. Guideline panels should strive for the sort of explicit approach that this example represents.

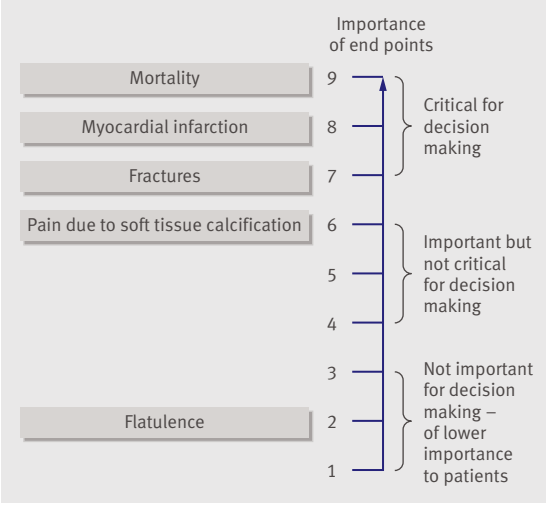

Fig 1 Hierarchy of outcomes according to importance to patients to assess effect of phosphate lowering drugs in patients with renal failure and hyperphosphataemia

Gordon H Guyatt professor, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada L8N 3Z5

Andrew D Oxman researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs Plass, 0130 Oslo, Norway

Gunn E Vist researcher, Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs Plass, 0130 Oslo, Norway

Regina Kunz associate professor, Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, 4031 Basel, Switzerland

Yngve Falck-Ytter assistant professor, Division of Gastroenterology, Case Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA

Holger J Schünemann professor, Department of Epidemiology, Italian National Cancer Institute Regina Elena, Rome, Italy

for the GRADE Working Group

Correspondence to:
G H Guyatt, CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, Room 2C12, 1200 Main Street, West Hamilton, ON, Canada L8N 3Z5
guyatt@mcmaster.ca

This is a series of five articles that explain the GRADE system for rating the quality of evidence and strength of recommendations

Judging the quality of evidence requires consideration of the context

GRADE provides a definition for the quality of evidence in the context of making recommendations. The quality of evidence reflects the extent to which confidence in an estimate of the effect is adequate to support recommendations. This definition has two important implications. Firstly, guideline panels must make judgments about the quality of evidence relative to the specific context in which they are using the evidence. Secondly, because systematic reviews do not—or at least should not—make recommendations, they require a different definition. For systematic reviews, the quality of evidence reflects the extent of confidence that an estimate of effect is correct.

Study design is important in determining the quality of evidence

As with early systems of grading the quality of evidence,[4] GRADE's approach begins with the study design. For recommendations addressing alternative management strategies—as opposed to issues of establishing prognosis or the accuracy of diagnostic tests—randomised trials provide, in general, stronger evidence than do observational studies. Rigorous observational studies provide stronger evidence than uncontrolled case series. In the GRADE approach to quality of evidence, randomised trials without important limitations constitute high quality evidence. Observational studies without special strengths or important limitations constitute low quality evidence.

Five limitations can reduce the quality of the evidence

The GRADE approach involves making separate ratings for quality of evidence for each patient important outcome and identifies five factors that can lower the quality of the evidence (see box).[5] These factors can

downgrade the quality of observational studies as well as randomised controlled trials.

Study limitations

Confidence in recommendations decreases if studies have major limitations that may bias their estimates of the treatment effect.[6] These limitations include lack of allocation concealment; lack of blinding—particularly if outcomes are subjective and their assessment highly susceptible to bias; large losses to follow-up; failure to adhere to an intention to treat analysis; stopping early for benefit[7]; or failure to report outcomes (typically those for which no effect was observed).

For example, most of the randomised trials examining the relative impact of standard wide tumour resection compared with a modified Whipple procedure for pancreatic carcinoma were limited by lack of optimal concealment, lack of possible blinding of patients and adjudicators of outcome, and substantial losses to follow-up. Thus the quality of evidence for each of the important outcomes was no higher than moderate (table 1).

Inconsistent results

Widely differing estimates of the treatment effect (heterogeneity or variability in results) across studies suggest true differences in underlying treatment effect. Variability may arise from differences in populations (for example, drugs may have larger relative effects in sicker populations), interventions (for example, larger effects with higher drug doses), or outcomes (for example, diminishing treatment effect with time). When heterogeneity exists but investigators fail to identify a plausible explanation, the quality of evidence decreases.

For example, the randomised trials of alternative approaches to the Whipple procedure yielded widely differing estimates of effects on gastric emptying, thus further decreasing the quality of the evidence (fig 2).

Factors in deciding on quality of evidence

Factors that might decrease quality of evidence
- Study limitations
- Inconsistency of results
- Indirectness of evidence
- Imprecision
- Publication bias

Factors that might increase quality of evidence
- Large magnitude of effect
- Plausible confounding, which would reduce a demonstrated effect
- Dose-response gradient

Indirectness of evidence

Guideline developers face two types of indirectness of evidence. The first occurs when considering, for example, use of one of two active drugs. Although randomised comparisons of the drugs may be unavailable, randomised trials may have compared one drug with placebo and the other with placebo. Such trials allow

Table 1 | GRADE evidence profile for impact of surgical alternatives for pancreatic cancer from systematic review and meta-analysis of randomised controlled trials in inpatient hospitals of pylorus preserving versus standard Whipple pancreaticoduodenectomy for pancreatic or periampullary cancer by Karanicolas et al[11]

	Quality assessment					Summary of findings			
No of studies (No of participants)	Study limitations*	Consistency	Directness	Precision	Publication bias	Relative effect† (95% CI)	Best estimate of Whipple group risk	Absolute effect (95% CI)	Quality
Five year mortality:									
3 (229)	Serious limitations (−1)	No important inconsistency	Direct	No important imprecision	Unlikely	0.98 (0.87 to 1.11)	82.5%	20 less/1000; 120 less to 80 more	+++, moderate
In-hospital mortality:									
6 (490)	Serious limitations (−1)	No important inconsistency	Direct	Imprecision (−1)‡	Unlikely	0.40 (0.14 to 1.13)	4.9%	20 less/1000; (50 less to 10 more)	++, low
Blood transfusions (units):									
5 (320)	Serious limitations (−1)	No important inconsistency	Direct	No important imprecision	Unlikely	—	2.45 units	−0.66 (−1.06 to −0.25); favours pylorus preservation	+++, moderate
Biliary leaks:									
3 (268)	Serious limitations (−1)	No important inconsistency	Direct	Imprecision (−1)‡	Unlikely	4.77 (0.23 to 97.96)	0	20 more/1000 20 less to 50 more	++, low
Hospital stay (days):									
5 (446)	Serious limitations (−1)	No important inconsistency	Direct	Imprecision (−1)‡	Unlikely	—	19.17 days	−1.45 (−3.28 to 0.38); favours pylorus preservation	++, low
Delayed gastric emptying:									
5 (442)	Serious limitations (−1)	Unexplained heterogeneity (−1)§	Direct	Imprecision (−1)‡	Unlikely	1.52 (0.74 to 3.14)	25.5%	110 more/1000; 80 less to 290 more	+, very low

*Unclear allocation concealment in all studies, patients blinded in only one study, outcome assessors not blinded in any study, >20% loss to follow-up in three studies, not analysed using intention to treat in one study.
†Relative risks (95% confidence intervals) are based on random effect models.
‡Confidence interval includes possible benefit from both surgical approaches.
§I2=72.6%, P=0.006.

indirect comparisons of the magnitude of effect of both drugs. Such evidence is of lower quality than would be provided by head to head comparisons of the drugs.

The second type of indirectness of evidence includes differences between the population, intervention, comparator to the intervention, and outcome of interest, and those included in the relevant studies. Table 2 presents examples of each.

Imprecision

When studies include relatively few patients and few events and thus have wide confidence intervals, a guideline panel will judge the quality of the evidence to be lower. For example, most of the outcomes for alternatives to the Whipple procedure include both important effects and no effects at all, and some include important differences in both directions (table 1).

Publication bias

The quality of evidence will be reduced if investigators fail to report studies (typically those that show no effect). A prototypical situation that should elicit suspicion of reporting bias occurs when published evidence is limited to a small number of trials, all of which are funded by industry.

Three factors can increase the quality of evidence

Although well done observational studies generally yield low quality evidence, in unusual circumstances they may produce moderate or even high quality evidence (see box).[8]

When methodologically strong observational studies yield large or very large and consistent estimates of the magnitude of a treatment effect, we may be confident about the results. In those situations, although the observational studies are likely to provide an overestimate of the true effect, the weak study design is unlikely to explain all of the apparent benefit.

The larger the magnitude of effect, the stronger becomes the evidence. For example, a meta-analysis of observational studies showed that bicycle helmets reduce the risk of head injuries in cyclists involved in a crash by a large margin (odds ratio 0.31, 95% confidence interval 0.26 to 0.37).[9] This large effect suggests a rating of moderate quality evidence. A meta-analysis of observational studies evaluating the impact of warfarin prophylaxis in cardiac valve replacement found that the relative risk for thromboembolism with warfarin was 0.17 (95% confidence interval 0.13 to 0.24). This very large effect suggests a rating of high quality evidence.

The presence of a dose-response gradient, or a situation in which all plausible biases would decrease the magnitude of effect, also increase the quality of the evidence.

Critical outcomes determine the rating of quality of evidence across outcomes

Recommendations depend on evidence for several patient important outcomes and the quality of evidence for each of those outcomes. How should the quality of evidence be rated across outcomes if quality differs? This occurred in the Whipple procedure example in which the evidence varied from moderate to very low quality (table 1).

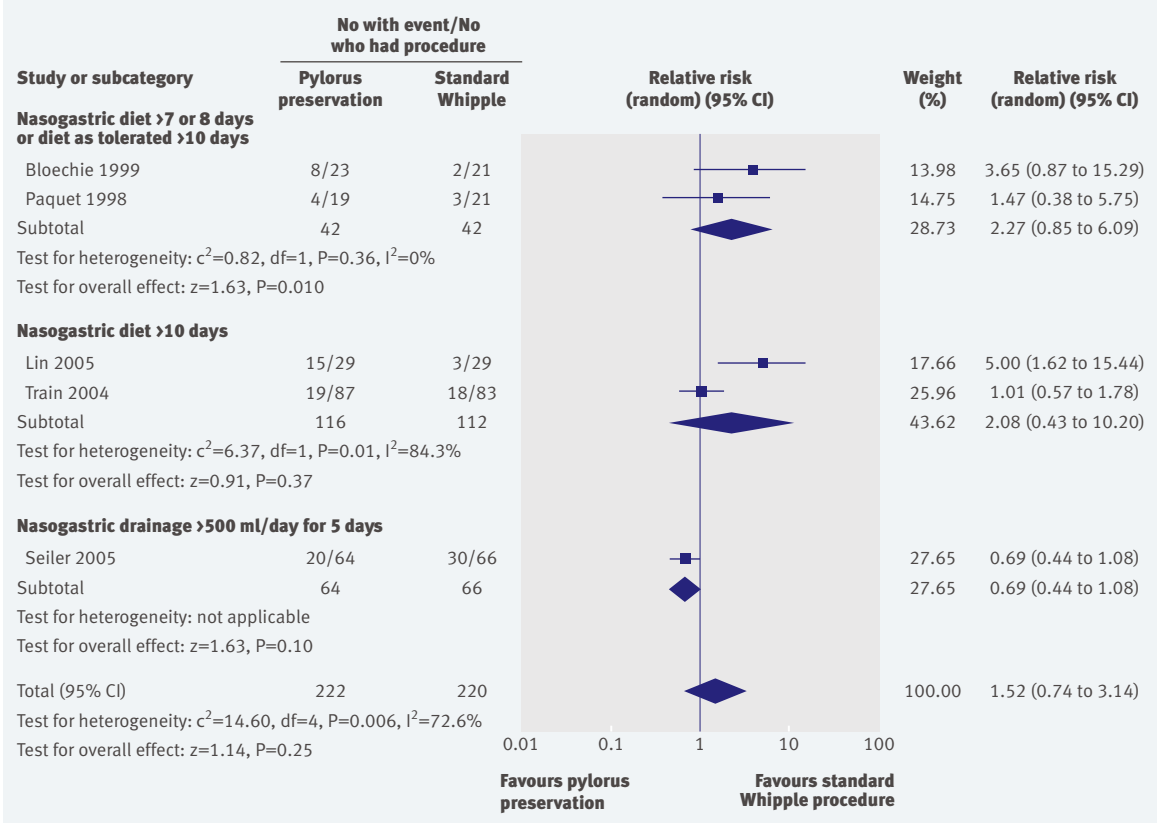

Fig 2 Effect on delayed gastric emptying of pylorus preserving pancreaticoduodenectomy compared with standard Whipple procedure for pancreatic adenocarcinoma

Table 2 | **Quality of evidence is weaker if comparisons in trials are indirect**

Question of interest	Source of indirectness
Relative effectiveness of alendronate and risedronate in osteoporosis	Indirect comparison: randomised trials have compared alendronate with placebo and risedronate with placebo, but trials comparing alendronate with risedronate are unavailable
Oseltamivir for prophylaxis of avian flu caused by influenza A (H5N1) virus	Differences in population: randomised trials of oseltamivir are available for seasonal influenza, but not for avian flu
Sigmoidoscopic screening for prevention of mortality from colon cancer	Differences in intervention: randomised trials of faecal occult blood screening provide indirect evidence, bearing on potential effectiveness of sigmoidoscopy
Choice of drug for schizophrenia	Differences in comparator: series of trials comparing newer generation neuroleptic agents with fixed doses of haloperidol 20 mg provide indirect evidence of how newer agents would compare with lower, flexible doses of haloperidol that clinicians typically use
Rosiglitazone for prevention of diabetic complications in patients at high risk of diabetes	Differences in outcome: randomised trial shows delay in development of biochemical diabetes with rosiglitazone but was underpowered to tackle diabetic complications

The GRADE approach suggests that guideline developers should consider the quality of evidence across outcomes as that associated with the critical outcome with the lowest quality evidence. Thus for the Whipple procedure example, if those making recommendations thought that gastric emptying problems were crucial, the rating for quality of evidence across outcomes would be very low. If gastric emptying was important but not critical, the quality rating across outcomes would be low (on the basis of results from the clearly critical perioperative mortality) despite the presence of moderate quality evidence on survival at five years (table 1).

Evidence profiles provide simple, transparent summaries

Busy clinicians require succinct, transparent, easily digested summaries on evidence. The GRADE process facilitates the creation of summaries, such as in table 2, which presents the relative effect of standard versus more limited resection for patients with pancreatic carcinoma.

Conclusion

GRADE provides a clearly articulated and comprehensive methodology for rating and summarising the quality of evidence supporting management recommendations. Although judgments will always be required for each step, the systematic and transparent GRADE approach allows scrutiny of and debate about those judgments.

SUMMARY POINTS

A guideline's formulation should include a clear question with specification of all outcomes of importance to patients

GRADE offers four levels of evidence quality: high, moderate, low, and very low

Randomised trials begin as high quality evidence and observational studies as low quality evidence

Quality may be downgraded as a result of limitations in study design or implementation, imprecision of estimates (wide confidence intervals), variability in results, indirectness of evidence, or publication bias

Quality may be upgraded because of a very large magnitude of effect, a dose-response gradient, and if all plausible biases would reduce an apparent treatment effect

Critical outcomes determine the overall quality of evidence

Evidence profiles provide simple, transparent summaries

Contributors: All authors, including members of the GRADE Working Group, contributed to the development of the ideas in the manuscript and read and approved the manuscript. GHG wrote the first draft and collated comments from authors and reviewers for subsequent iterations. He is guarantor for this paper. All authors listed in the byline contributed ideas about structure and content, provided examples, reviewed drafts of the manuscript, and provided feedback.

The members of the GRADE Working Group are Phil Alderson, Pablo Alonso-Coello, Jeff Andrews, David Atkins, Hilda Bastian, Hans de Beer, Jan Brozek, Francoise Cluzeau, Jonathan Craig, Ben Djulbegovic, Yngve Falck-Ytter, Beatrice Fervers, Signe Flottorp, Paul Glasziou, Gordon H Guyatt, Margaret Haugh, Robin Harbour, Mark Helfand, Sue Hill, Roman Jaeschke, Katharine Jones, Ilkka Kunnamo, Regina Kunz, Alessandro Liberati, Merce Marzo, James Mason, Jacek Mrukowics, Susan Norris, Andrew D Oxman, Vivian Robinson, Holger J Schünemann, Tessa Tan Torres, David Tovey, Peter Tugwell, Mariska Tuut, Helena Varonen, Gunn E Vist, Craig Wittington, John Williams, and James Woodcock.

Funding: No specific funding.

Competing interests: All authors are involved in the dissemination of GRADE, and GRADE's success has a positive influence on their academic career. Authors listed in the byline have received travel reimbursement and honorariums for presentations that included a review of GRADE's approach to rating quality of evidence and grading recommendations. GHG acts as a consultant to UpToDate; his work includes helping UpToDate in their use of GRADE. HJS is documents editor and methodologist for the American Thoracic Society; one of his roles in these positions is helping implement the use of GRADE. He is supported by "The human factor, mobility and Marie Curie actions scientist reintegration European Commission grant: IGR 42192—GRADE."

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. *BMJ* 2004;328:1490.
2 Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988;138:697-703.
3 Schunemann H, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 10. Integrating values and consumer involvement. *Health Res Policy Syst* 2006;5:4-22.
4 Fletcher SW, Spitzer WO. Approach of the Canadian Task Force to the periodic health examination. *Ann Intern Med* 1980;92(2 Pt 1):253-4.
5 Schunemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, et al. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *Am J Respir Crit Care Med* 2006;174:605-14.
6 Guyatt G, Cook D, Devereaux PJ, Meade M, Straus S. Therapy. In: Guyatt G, Rennie D, eds. *The users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: AMA publications, 2002.
7 Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA* 2005;294:2203-9.
8 Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007;334:349-51.
9 Thompson DC, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. *Cochrane Database Syst Rev* 2000;(2):CD001855.
10 Cannegieter SC, Rosendaal FR, Briet E. Thromboembolic and bleeding complications in patients with mechanical heart valve prostheses. *Circulation* 1994;89:635-41.
11 Karanicolas PJ, Davies E, Kunz R, Briel M, Koka HP, Payne DM, et al. The pylorus: take it or leave it? Systematic review and meta-analysis of pylorus-preserving versus standard whipple pancreaticoduodenectomy for pancreatic or periampullary cancer. *Ann Surg Oncol* 2007;14:1825-34.