Grading quality of evidence and strength of recommendations

Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) Working Group

Clinical guidelines are only as good as the evidence and judgments they are based on. The GRADE approach aims to make it easier for users to assess the judgments behind recommendations

Correspondence to:
Andrew D Oxman, Informed Choice Research Department, Norwegian Health Services, PO Box 7004, St Olavs plass, 0130 Oslo, Norway
oxman@online.no

BMJ 2004;328:1490–4

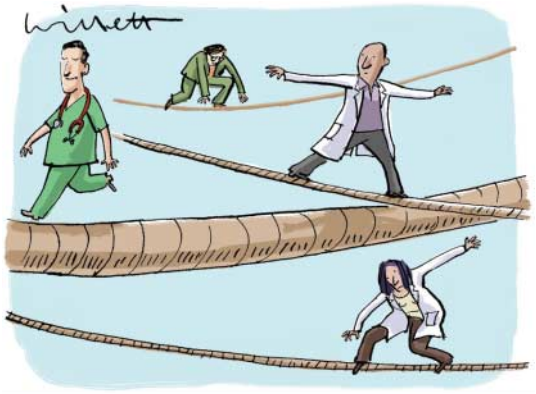
Healthcare workers using clinical practice guidelines and other recommendations need to know how much confidence they can place in the recommendations. Systematic and explicit methods of making judgments can reduce errors and improve communication. We have developed a system for grading the quality of evidence and the strength of recommendations that can be applied across a wide range of interventions and contexts. In this article we present a summary of our approach from the perspective of users of guidelines.

What makes a good guideline?

Judgments about evidence and recommendations are complex. Consider, for example, the choice between selective serotonin reuptake inhibitors and tricyclic antidepressants for the treatment of moderate depression. Clinicians must decide which outcomes to consider, which evidence to include for each outcome, how to assess the quality of that evidence, and how to determine if selective serotonin reuptake inhibitors do more good than harm compared with tricyclics. Because resources are always limited and money that is spent on serotonin reuptake inhibitors cannot be used elsewhere, they may also need to decide whether any incremental health benefits are worth the additional costs.

It is not practical for individual clinicians and patients to make unaided judgments for each clinical decision. Clinicians and patients commonly use clinical practice guidelines as a source of support. Users of guidelines need to know how much confidence they can place in the evidence and recommendations. We describe the factors on which our confidence should be based and a systematic approach for making the complex judgments that go into clinical practice guidelines, either implicitly or explicitly. To achieve simplicity in our presentation we do not discuss all the nuances, some of which are discussed in the longer version of this article on bmj.com.

The GRADE Working Group began as an informal collaboration of people with an interest in tackling the shortcomings of present grading systems. Table 1 summarises these shortcomings and the ways in which we have overcome them. The GRADE system enables more consistent judgments, and communication of



such judgments can support better-informed choices in health care. Box 1 shows the steps in developing and implementing guidelines from prioritising problems through evaluating their implementation. We focus here on grading the quality of evidence and strength of recommendations.

Definitions

We have used the following definitions: the quality of evidence indicates the extent to which we can be confident that an estimate of effect is correct; the strength of a recommendation indicates the extent to which we can be confident that adherence to the recommendation will do more good than harm.

The steps in our approach are to make sequential judgments about:

- The quality of evidence across studies for each important outcome
- Which outcomes are critical to a decision
- The overall quality of evidence across these critical outcomes
- The balance between benefits and harms
- The strength of recommendations

All of these judgments depend on having a clearly defined question and considering all of the outcomes that are likely to be important to those affected. The question should identify which options are being com-

ELPS *This is an abridged version; the full version is on bmj.com*

Box 1: Sequential process for developing guidelines

First steps
1. *Establishing the process*—For example, prioritising problems, selecting a panel, declaring conflicts of interest, and agreeing on group processes

Preparatory steps
2. *Systematic review*—The first step is to identify and critically appraise or prepare systematic reviews of the best available evidence for all important outcomes
3. *Prepare evidence profile for important outcomes*—Profiles are needed for each subpopulation or risk group, based on the results of systematic review, and should include a quality assessment and a summary of findings

Grading quality of evidence and strength of recommendations
4. *Quality of evidence for each outcome*—Judged on information summarised in the evidence profile and based on the criteria in table 2
5. *Relative importance of outcomes*—Only important outcomes should be included in evidence profiles. The included outcomes should be classified as critical or important (but not critical) to a decision
6. *Overall quality of evidence*—The overall quality of evidence should be judged across outcomes based on the lowest quality of evidence for any of the critical outcomes.
7. *Balance of benefits and harms*—The balance of benefits and harms should be classified as net benefits, trade-offs, uncertain trade-offs, or no net benefits based on the important health benefits and harms
8. *Balance of net benefits and costs*—Are incremental health benefits worth the costs? Because resources are always limited, it is important to consider costs (resource utilisation) when making a recommendation
9. *Strength of recommendation*—Recommendations should be formulated to reflect their strength—that is, the extent to which one can be confident that adherence will do more good than harm

Subsequent steps
10. *Implementation and evaluation*—For example, using effective implementation strategies that address barriers to change, evaluation of implementation, and keeping up to date

pared (for example, selective serotonin reuptake inhibitors and tricyclic antidepressants), for whom (moderately depressed adult patients), and in what setting (primary care in England).

Quality of evidence

Judgments about quality of evidence should be guided by a systematic review of available evidence. Reviewers should consider four key elements: study design, study quality, consistency, and directness (box 2). Study design refers to the basic study design, which we have broadly categorised as observational studies and randomised trials. Study quality refers to the detailed study methods and execution. Consistency refers to the similarity of estimates of effect across studies. Directness refers to the extent to which the people, interventions, and outcome measures are similar to those of interest. Another type of indirect evidence arises when there are no direct comparisons of interventions and investigators must make comparisons across studies.

The quality of evidence for each main outcome can be determined after considering each of these four elements. Our approach initially categorises evidence based on study design into randomised trials and observational studies (box 2). We then suggest considering whether the studies have serious limitations, important inconsistencies in the results, or whether uncertainty about the directness of the evidence is warranted.

Additional considerations that can lower the quality of evidence include imprecise or sparse data and a high risk of reporting bias. Additional considerations that can raise the quality of evidence include a very strong association (for example, a 50-fold risk of poisoning fatalities with tricyclic antidepressants; see table 2) or strong association (for example, a threefold increased risk of head injuries among cyclists who do not use helmets compared with those who do[1]) and evidence of a dose-response gradient. Box 3 gives our suggested definitions for grading the quality of the evidence.

The same rules should be applied to judgments about the quality of evidence for harms and benefits. Important plausible harms can and should be included in evidence summaries by considering the indirect evidence that makes them plausible. For example, if there is concern about anxiety in relation to screening for melanoma and no direct evidence is found, it may be appropriate to consider evidence from studies of other types of screening.

Judgments about the quality of evidence for important outcomes across studies can and should be made in the context of systematic reviews, such as Cochrane reviews. Judgments about the overall quality of evidence, trade-offs, and recommendations typically require information beyond the results of a review.

Other systems have commonly based judgments of the overall quality of evidence on the quality of evidence for the benefits of interventions. When the risk of an adverse effect is critical for a judgment, and evidence regarding that risk is weaker than evidence of benefit, ignoring uncertainty about the risk of harm is problematic. We suggest that the lowest quality of evidence for any of the outcomes that are critical to making a decision should provide the basis for rating overall quality of evidence.

Recommendations

Does the intervention do more good than harm?
Recommendations involve a trade-off between benefits and harms. Making that trade-off inevitably involves placing, implicitly or explicitly, a relative value on each outcome. We suggest making explicit judgments about the balance between the main health benefits and harms before considering costs. Does the intervention do more good than harm?

Recommendations must apply to specific settings and particular groups of patients whenever the benefits and harms differ across settings or patient groups. For instance, consider whether you should recommend that patients with atrial fibrillation receive warfarin to reduce their risk of stroke, despite the increase in bleeding risk that will result. Recommendations, or their strength, are likely to differ in settings where regular monitoring of the intensity of anti-

Table 1 Comparison of GRADE and other systems

Factor	Other systems	GRADE	Advantages of GRADE system*
Definitions	Implicit definitions of quality (level) of evidence and strength of recommendation	Explicit definitions	Makes clear what grades indicate and what should be considered in making these judgments
Judgments	Implicit 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	Sequential, explicit judgments	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	Not considered for each important outcome. Judgments about quality of evidence are often based on study design alone	Systematic and explicit consideration of study design, study quality, consistency, and directness of evidence in judgments about quality of evidence	Ensures these factors are considered appropriately
Other factors that can affect quality of evidence	Not explicitly taken into account	Explicit consideration of imprecise or sparse data, reporting bias, strength of association, evidence of a dose-response gradient, and plausible confounding	Ensures consideration of other factors
Overall quality of evidence	Implicitly based on the quality of evidence for benefits	Based on the lowest quality of evidence for any of the outcomes that are critical to making a decision	Reduces likelihood of mislabelling overall quality of evidence when evidence for a critical outcome is lacking
Relative importance of outcomes	Considered implicitly	Explicit judgments about which outcomes are critical, which ones are important but not critical, and which ones are unimportant and can be ignored	Ensures appropriate consideration of each outcome when grading overall quality of evidence and strength of recommendations
Balance between health benefits and harms	Not explicitly considered	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
Whether incremental health benefits are worth the costs	Not explicitly considered	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	Inconsistent presentation	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	Seldom used by more than one organisation and little, if any empirical evaluation	International collaboration across wide range of organisations in development and evaluation	Builds on previous experience to achieve a system that is more sensible, reliable, and widely applicable

*Most other approaches do not include any of these advantages, although some may incorporate some of these advantages.

coagulation is available and settings where it is not. Furthermore, recommendations (or their strength) are likely to differ in patients at low risk of stroke (those under 65 without any comorbidity) and patients at higher risk (such as older patients with heart failure) because of differences in the absolute reduction in risk. Recommendations must therefore be specific to a patient group and a practice setting.

Box 2: Criteria for assigning grade of evidence

Type of evidence
Randomised trial = high
Observational study = low
Any other evidence = very low

Decrease grade if:

- Serious (−1) or very serious (−2) limitation to study quality
- Important inconsistency (−1)
- Some (−1) or major (−2) uncertainty about directness
- Imprecise or sparse data (−1)
- High probability of reporting bias (−1)

Increase grade if:

- Strong evidence of association—significant relative risk of >2 (<0.5) based on consistent evidence from two or more observational studies, with no plausible confounders (+1)
- Very strong evidence of association—significant relative risk of >5 (<0.2) based on direct evidence with no major threats to validity (+2)
- Evidence of a dose response gradient (+1)
- All plausible confounders would have reduced the effect (+1)

Those making a recommendation should consider four main factors:

- The trade-offs, taking into account the estimated size of the effect for the main outcomes, the confidence limits around those estimates, and the relative value placed on each outcome
- The quality of the evidence
- Translation of the evidence into practice in a specific setting, taking into consideration important factors that could be expected to modify the size of the expected effects, such as proximity to a hospital or availability of necessary expertise
- Uncertainty about baseline risk for the population of interest.

If there is uncertainty about translating the evidence into practice in a specific setting, or uncertainty about baseline risk, this may lower our confidence in a recommendation. For example, if an intervention has serious adverse effects as well as important benefits, a recommendation is likely to be much less certain when the baseline risk of the population of interest is uncertain than when it is known.

Box 3: Definitions of grades of evidence

High = Further research is 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.

Table 2 Quality assessment of trials comparing selective serotonin reuptake inhibitors (SSRIs) with tricyclic antidepressants for treatment of moderate depression in primary care[2]

	Quality assessment					Summary of findings					
						No of patients		Effect			
No of studies	Design	Quality	Consistency	Directness	Other modifying factors*	SSRIs	Tricyclics	Relative (95% CI)	Absolute	Quality	Importance
Depression severity (measured with Hamilton depression rating scale after 4 to 12 weeks)											
Citalopram (8) Fluoxetine (38) Fluvoxamine (25) Nefazodone (2) Paroxetine (18) Sertraline (4) Venlafaxine (4)	Randomised controlled trials	No serious limitations	No important inconsistency	Some uncertainty about directness (outcome measure)†	None	5044	4510	WMD 0.034 (–0.007 to 0.075)	No difference	Moderate	Critical
Transient side effects resulting in discontinuation of treatment											
Citalopram (8) Fluoxetine (50) Fluvoxamine (27) Nefazodone (4) Paroxetine (23) Sertraline (6) Venlafaxine (5)	Randomised controlled trials	No serious limitations	No important inconsistency	Direct	None	1948/7032 (28%)	2072/6334 (33%)	RRR 13% (5% to 20%)	5/100	High	Critical
Poisoning fatalities§											
UK Office for National Statistics (1)	Observational data	Serious limitation‡	Only one study	Direct	Very strong association	1/100 000/ year of treatment	58/100 000/ year of treatment	RRR 98% (97% to 99%)§	6/10 000	Moderate	Critical

WMD = weighted mean difference, RRR = relative risk reduction.
*Imprecise or sparse data, a strong or very strong association, high risk of reporting bias, evidence of a dose-response gradient, effect of plausible residual confounding.
†There was uncertainty about the directness of the outcome measure because of the short duration of the trials.
‡It is possible that people at lower risk were more likely to have been given SSRIs and it is uncertain if changing antidepressant would have deterred suicide attempts.
§There is uncertainty about the baseline risk for poisoning fatalities.

We suggest using the following categories for recommendations:

"Do it" or *"don't do it"*—indicating a judgment that most well informed people would make;

"Probably do it" or *"probably don't do it"*—indicating a judgment that a majority of well informed people would make but a substantial minority would not.

A recommendation to use or withhold an intervention does not mean that all patients should be treated identically. Nor does it mean that clinicians should not involve patients in the decision, or explain the merits of the alternatives. However, because most well informed patients will make the same choice, the explanation of the relative merits of the alternatives may be relatively brief. A recommendation is intended to facilitate an appropriate decision for an individual patient or a population. It should therefore reflect what people would likely choose, based on the evidence and their own values or preferences in relation to the expected outcomes. A recommendation to probably do something indicates a need for clinicians to consider patients' values and preferences more carefully when offering them the intervention.

In some instances it may not be appropriate to make a recommendation because of unclear trade-offs or lack of agreement. When this is due to a lack of good quality evidence, specific research should be recommended that would provide the evidence that is needed to inform a recommendation.

Are the incremental health benefits worth the costs?

Because spending money on one intervention means less money to spend on another, recommendations implicitly (if not explicitly) rely on judgments about the value of the incremental health benefits in relation to the incremental costs. Costs—the monetary value of resources used—are important considerations in making recommendations, but they are context specific, change over time, and their magnitude may be difficult to estimate. While recognising the difficulty of accurate estimating costs, we suggest that the incremental costs of healthcare alternatives should be considered explicitly alongside the expected health benefits and harms. When relevant and available, disaggregated costs (differences in use of resources) should be presented in evidence profiles along with important outcomes. The

Summary points

Organisations have used various systems to grade the quality of evidence and strength of recommendations

Differences and shortcomings in these grading systems can be confusing and impede effective communication

A systematic and explicit approach to making judgments about the quality of evidence and the strength of recommendations is presented

The approach takes into account study design, study quality, consistency, and directness in judging the quality of evidence for each important outcome

The balance between benefits and harms, quality of evidence, applicability, and the certainty of the baseline risk are all considered in judgments about the strength of recommendations

quality of the evidence for differences in use of resources should be graded by using the approach outlined above for other important outcomes.

How it works in practice

Table 2 shows an example of the system applied to evidence from a systematic review comparing selective serotonin reuptake inhibitors with tricyclic antidepressants conducted in 1997.[2] After discussion, we agreed that there was moderate quality evidence for the relative effects of both types of drugs on severity of depression and poisoning fatalities and high quality evidence for transient side effects. We then reached agreement that the overall quality of evidence was moderate and that there were net benefits in favour of serotonin reuptake inhibitors (no difference in severity of depression, fewer transient side effects, and fewer poisoning fatalities). Although we agreed that there seemed to be net benefits, we concluded with a recommendation to "probably" use serotonin reuptake inhibitors because of uncertainty about the quality of the evidence. We had no evidence on relative costs in this exercise. Had we considered costs, this recommendation might have changed.

Conclusions

We have attempted to find a balance between simplicity and clarity in our system for grading the quality of evidence and strength of recommendations. Regardless of how simple or complex a system is, judgments are always required. Our system provides a framework for structured reflection and can help to ensure that appropriate judgments are made, but it does not remove the need for judgment.

Contributors and sources: see bmj.com

Competing interests: Most of the members of the GRADE Working Group have a vested interest in another system of grading the quality of evidence and the strength of recommendations.

1 Thompson DC, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. *Cochrane Database Syst Rev* 2000;(2):CD001855.
2 North of England Evidence Based Guideline Development Project. *Evidence based clinical practice guideline: the choice of antidepressants for depression in primary care.* Newcastle upon Tyne: Centre for Health Services Research, 1997.

(Accepted 5 March 2004)

Medical researchers' ancillary clinical care responsibilities

Leah Belsky, Henry S Richardson

Investigation of participants in clinical trials may identify conditions unrelated to the study. Researchers need guidance on whether they have a duty to treat such conditions

Department of Clinical Bioethics, National Institutes of Health, 10 Center Drive, Bethesda, MD 20898, USA
Leah Belsky *fellow*
Henry S Richardson *visiting scholar*
Correspondence to: L Belsky lbelsky@mail.cc.nih.gov

BMJ 2004;328:1494–6

Researchers testing a new treatment for tuberculosis in a developing country discover some patients have HIV infection. Do they have a responsibility to provide antiretroviral drugs? In general, when do researchers have a responsibility to provide clinical care to participants that is not stipulated in the trial's protocol? This question arises regularly, especially in developing countries, yet (with rare exceptions[1]) existing literature and guidelines on research ethics do not consider ancillary clinical care. We propose an ethical framework that will help delineate researchers' responsibilities.

What is ancillary care?

Ancillary care is that which is not required to make a study scientifically valid, to ensure a trial's safety, or to redress research injuries. Thus, stabilising patients to enrol them in a research protocol, monitoring drug interactions, or treating adverse reactions to experimental drugs are not ancillary care. By contrast, following up on diagnoses found by protocol tests or treating ailments that are unrelated to the study's aims would be ancillary care.

Two extreme views

When asked how much ancillary care they should provide to participants, the first reaction of many clinical researchers, especially those working in developing countries, is that they must provide whatever ancillary care their participants need. From an ethical perspective, this response makes sense. Research participants in trials in the developing world are typically desperately poor and ill, and everyone arguably has a duty to rescue those in need, at least when they can do so at minimal cost to themselves.[2 3] Yet this response fails to acknowledge that the goal of research is to generate knowledge not care for patients.[4 5] When researchers consider that offering ancillary care this broadly may drain limited human and financial resources and confound study results, they tend to retreat from this position.

Some researchers veer to the opposite extreme. "We may be doctors," they note, "but these are our research participants, not our patients, so we owe them nothing beyond what is needed to complete the study safely and successfully—that is, we owe them no ancillary care." But this extreme position is ethically questionable. Consider the case of researchers studying a rare disease. It is ethically unacceptable to say to a participant, "We are going to monitor the toxicity and effectiveness of this experimental drug, and we will make sure it does not kill you, but we are not going to provide any palliative care for your condition." Closely monitoring a participant's disease without being willing to treat it in any way amounts to treating him or her as a mere means to the end of research.