Education and debate

Grading quality of evidence and strength of recommendations

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

Summary

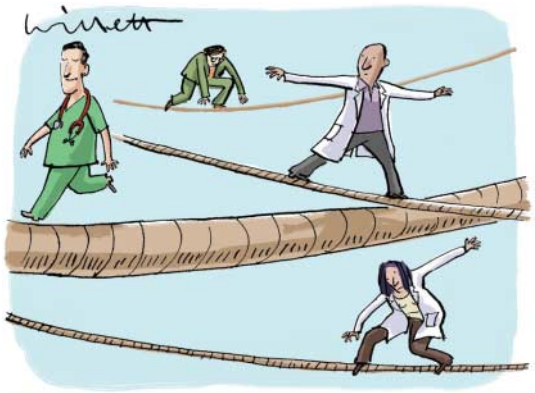
Users of clinical practice guidelines and other recommendations need to know how much confidence they can place in the recommendations. Systematic and explicit methods of making judgments can reduce errors and improve communication. We have developed a system for grading the quality of evidence and the 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 a guideline user. Judgments about the strength of a recommendation require consideration of the balance between benefits and harms, the quality of the evidence, translation of the evidence into specific circumstances, and the certainty of the baseline risk. It is also important to consider costs (resource utilisation) before making a recommendation. Inconsistencies among systems for grading the quality of evidence and the strength of recommendations reduce their potential to facilitate critical appraisal and improve communication of these judgments. Our system for guiding these complex judgments balances the need for simplicity with the need for full and transparent consideration of all important issues.

Introduction

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 selective 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 these judgments unaided for each clinical decision. Clinicians and patients commonly use clinical practice guidelines as a source of support—that is, recommendations that have been systematically developed by panels of people with access to the available evidence, an understanding of the clinical problem and research methods, and sufficient time for reflection.

Users of systematically developed guidelines need to know how much confidence they can place in 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 or other healthcare recommendations, either implicitly or explicitly. To achieve simplicity in our presentation we do not discuss all the nuances or provide detailed guidance that guideline panels would need to apply our approach. This can be obtained from the authors (www.GradeWorkingGroup.org).



A systematic and explicit approach to making judgments about the quality of evidence and the strength of recommendations can help to prevent errors, facilitate critical appraisal of these judgments, and can help to improve communication of this information. Since the 1970s a growing number of organisations have employed various systems to grade the quality (level) of evidence and the strength of recommendations.[1–28] Unfortunately, different organisations 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.

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.

Members of GRADE Working group are listed at the end of this article

Definitions

We have used the following definitions: 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.

Judgments about the quality of evidence require assessments of the validity of the results of individual studies for important outcomes. Explicit criteria should be used in making these judgments.[26 29–32] The steps in our approach, which follow these judgments, 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 compared (for example, selective serotonin reuptake inhibitors and tricyclic antidepressants), for whom (moderately depressed adult patients), and in what setting (primary care in England).

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 reviews, 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

Quality of evidence for each important outcome

A systematic review of available evidence should guide these judgments. Reviewers should consider four key elements: study design, study quality, consistency, and directness.

Study design

Study design refers to the basic study design, which we have broadly categorised as observational studies and randomised trials. Both logical arguments and empirical evidence support this.[33–36] Although observational studies commonly have results that are similar to those of randomised trials, this is not always the case. One dramatic example of such a discrepancy is the different results of observational studies that suggested hormone replacement therapy decreased the risk of coronary heart disease and subsequent randomised trials that found no reduction in risk and even an increased risk.[37 38] Unfortunately, it is not possible to know in advance whether observational studies accurately predict the findings of subsequent randomised trials. Once the results of high quality randomised trials are available, few people would argue for continuing to base recommendations on non-randomised studies with discrepant results.

On the other hand, randomised trials are not always feasible and, in some instances, observational studies may provide better evidence, as is generally the case for rare adverse effects. Moreover, the results of randomised trials may not always be applicable—for example, if the participants are highly selected and motivated relative to the population of interest. It is therefore essential to consider study quality, the consistency of results across studies, and the directness of the evidence, as well as the appropriateness of the study design. So, for example, well designed case series may provide high quality evidence for complication rates from surgery or procedures, such as intraoperative deaths or perforations after colonoscopy, which is more directly relevant than evidence from randomised trials. Similarly, cohort studies can provide high quality evidence for rates of recall or procedures precipitated by false positive screening results, such as biopsy rates after mammography.

Study quality

Study quality refers to the detailed study methods and execution. Reviewers should use appropriate criteria to assess study quality for each important outcome.[26 29–32] For randomised trials, for example, reviewers might use criteria such as the adequacy of allocation concealment, blinding, and follow up. Reviewers should make explicit their reasons for downgrading a quality rating. For example, they may state that failure to blind patients and physicians reduced the quality of evidence for an intervention's impact on pain severity and that they considered this a serious limitation.

Consistency

Consistency refers to the similarity of estimates of effect across studies. If there is important unexplained inconsistency in the results, our confidence in the estimate of effect for that outcome decreases. Differences in the direction of effect, the size of the differences in effect, and the significance of the differences guide the (inevitably somewhat arbitrary) decision about whether important inconsistency exists. Separate estimates of magnitude of effect for different subgroups should follow when investigators identify a compelling explanation for inconsistency. For

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.

instance, differences in the effect of carotid endarterectomy on high and lower grade stenoses should lead to separate estimates for these two subgroups.

Directness

Directness refers to the extent to which the people, interventions, and outcome measures are similar to those of interest. For example, there may be uncertainty about the directness of the evidence if the people of interest are older, sicker, or have more comorbidity than those in the studies.[39] To determine whether important uncertainty exists, we can ask whether there is a compelling reason to expect important differences in the size of the effect. Because many interventions have more or less the same relative effects across most patient groups, we should not apply overly stringent criteria in deciding whether evidence is direct. For some therapies—for example, behavioural interventions in which cultural differences are likely to be important—more stringent criteria may be appropriate.

Similarly, reviewers may identify uncertainty about the directness of evidence for drugs that differ from those in the studies but are within the same class. Similar issues arise for other types of interventions. For instance, can you generalise results to a less intense counselling intervention than that used in a study, or to an alternative surgical technique? These judgments can be difficult,[40] and it is important for investigators to explain the rationale for the conclusions that they draw.

On the other hand, studies using surrogate outcomes generally provide less direct evidence than those using outcomes that are important to people. It is therefore prudent to use much more stringent criteria when considering the directness of evidence for surrogate outcomes. Examples of indirect evidence based on surrogate outcomes that subsequent trials showed to be misleading include suppression of cardiac arrhythmia in patients who have had a myocardial infarction as a surrogate for mortality,[41] changes in lipoproteins as a surrogate for coronary heart disease,[37] and bone density in postmenopausal women as a surrogate for fracture reduction.[42]

The accuracy of a diagnostic test is also a surrogate for important outcomes that might be affected by accurate diagnosis, including improved health outcomes from appropriate treatment and reduced harms from false positive results. Different criteria must be used when considering study design for studies of diagnostic accuracy. However, consideration of the directness of evidence is based on how confident we are of the relation between being classified correctly (as a true positive or negative) or incorrectly (as a false positive or negative) and important consequences of this. For example, there is consistent evidence from well designed studies that there are fewer false negative results with non-contrast helical computed tomography than with intravenous pyelography in the diagnosis of suspected acute urolithiasis.[43] However, there is major uncertainty about whether this has important health consequences.[44] Because of this, the quality of this evidence could be considered low for making a recommendation.

Another type of indirect evidence arises when there are no direct comparisons of interventions and investigators must make comparisons across studies. For example, this would be the case if there were randomised trials that compared selective serotonin reuptake inhibitors with placebo and tricyclics with placebo, but no trials that compared selective serotonin reuptake inhibitors with tricyclics. Indirect comparisons always leave greater uncertainty than direct comparisons because of all the other differences between studies that can affect the results.[45]

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)[46]
- Very strong evidence of association—significant relative risk of >5 (<0.2) based on direct evidence with no major threats to validity (+2)[46]
- Evidence of a dose response gradient (+1)
- All plausible confounders would have reduced the effect (+1)

Combining the four components

The quality of evidence for each main outcome can be determined after considering each of the above elements: study design, study quality, consistency, and directness. Our approach initially categorises evidence based on study design into randomised trials and observational studies (cohort studies, case-control studies, interrupted time series analyses, and controlled before and after studies). 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 (box 2). We suggest the following definitions in grading the quality of the evidence:

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.

Limitations in study quality, important inconsistency of results, or uncertainty about the directness of the evidence can lower the grade of evidence. For instance, if all available studies have serious limitations, the grade will drop by one level, and if all studies have very serious limitations the grade will drop by two levels. Fatally flawed studies may be excluded.

Additional considerations that can lower the quality of evidence include imprecise or sparse data (box 3) and 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 compared with selective serotonin reuptake inhibitors, 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 that do[47])
- Evidence of a dose response gradient, or
- Presence of all plausible residual confounding would have reduced the observed effect. (For example, plausible explanatory factors that were not adjusted for in studies comparing mortality rates of for-profit and not-for-profit hospitals would have reduced the observed effect.[48] Thus, the evidence that for-profit hospitals have a higher risk of mortality is more convincing.)

These considerations act cumulatively. For example, if randomised trials have both serious limitations and there is uncertainty about the directness of the evidence, the grade of evidence would drop from high to low.

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.

Overall quality of evidence

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.

Outcomes that are important, but not critical, should be included in evidence profiles and should be considered when making judgments about the balance between health benefits and harms but should not be taken into consideration when grading the overall quality of evidence. Deciding whether an outcome is critical, important but not critical, or not important is a value judgment. So far as possible these judgments should take account of

Box 3: Imprecise or sparse data

There is not an empirical basis for defining imprecise or sparse data. Two possible definitions are:
- Data are sparse if the results include just a few events or observations and they are uninformative
- Data are imprecise if the confidence intervals are sufficiently wide that an estimate is consistent with either important harms or important benefits.

These different definitions can result in different judgments. Although it may not be possible to reconcile these differences, we offer the following guidance when considering whether to downgrade the quality of evidence due to imprecise or sparse data:
- The threshold for considering data imprecise or sparse should be lower when there is only one study. A single study with a small sample size (or few events) yielding wide confidence intervals spanning both the potential for harm and benefit should be considered as imprecise or sparse data
- Confidence intervals that are sufficiently wide that, irrespective of other outcomes, the estimate is consistent with conflicting recommendations should be considered as imprecise or sparse data

the values of those who will be affected by adherence to subsequent recommendations.

The decision regarding what is critical can be difficult. The plausibility of adverse outcomes may influence the decision regarding whether they are critical. Weak evidence about implausible putative harms should not lower the overall grade of evidence. Decisions about whether a putative harm is plausible may come from indirect evidence. For example, if there is important concern about serious adverse effects of a drug because of animal studies, the overall quality of evidence may receive a lower grade based on whatever human evidence is available for that particular adverse effect. Sometimes lack of evidence for plausible putative harms may make it impossible to assess the net benefit of an intervention. In these circumstances a guideline panel may elect to recommend additional research.

If the evidence for all of the critical outcomes favours the same alternative, and there is high quality evidence for some, but not all, of those outcomes, the overall quality of evidence might still be considered high. For example, there is high quality evidence that antiplatelet therapy reduces the risk of non-fatal stroke and non-fatal myocardial infarction in patients who have had a myocardial infarction. Although the evidence for all-cause mortality is of moderate quality, the overall quality of evidence might still be considered high, even if all cause mortality was considered a critical outcome.

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. It is often difficult to judge how much weight to give to different outcomes, and different people will often have different values. People making judgments on behalf of others are on stronger ground if they have evidence of the values of those affected. For instance, people making recommendations about chemotherapy for women with early breast cancer will be in a stronger position if they have evidence about the relative importance those women place on reducing the risk of a recurrence of breast cancer relative to avoiding the side effects of chemotherapy.

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 we 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 anticoagulation is available and settings where it is not. Furthermore, recommendations (or their strength) are likely to differ in patients at very 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. It is particularly important to consider the circumstances of disadvantaged populations when making recommendations and, when appropriate, modify recommendations to take into consideration differences between advantaged and disadvantaged populations.

We suggest using the following definitions to categorise the trade-offs:

Net benefits = the intervention clearly does more good than harm.

Trade-offs = there are important trade-offs between the benefits and harms.

Uncertain trade-offs = it is not clear whether the intervention does more good than harm.

No net benefits = the intervention clearly does not do more good than harm.

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

Box 4: Values are not right or wrong

The following example shows how different people might make different recommendations because of differences in values, even after agreeing on the evidence.
Question: Should the general population be screened for melanoma?
Setting: Primary care in the United States
Baseline risk: General population (melanoma incidence in 1995 was 13.3 per 100 000)
Reference: Helfand et al. *Screening for skin cancer. Systematic evidence review No 2.* Rockville, MD: Agency for Healthcare Research and Quality. April 2001. (AHRQ Publication No 01-S002.)
There is very low quality evidence for the accuracy of screening and for the outcome of lethal melanoma. Potential harms from screening include the consequences of false positive tests, but evidence regarding these is lacking. Based on this it is possible to conclude that the overall quality of evidence is very low and that there are uncertain net benefits from screening. Based on a single case-control study, the odds ratio for lethal melanoma was estimated to be 0.37 for screened versus not screened people. The lifetime risk of dying of melanoma was estimated to be 0.36% for white men.
Based on this evidence, many people might make a recommendation of "don't screen" because of placing a high value on avoiding the potential but unknown harms of screening healthy people relative to the uncertain benefits. However, some people might recommend "probably screen" because of placing a high value on the small but potentially important benefits of screening relative to the unknown potential harms. Under these circumstances, after taking into consideration costs, a panel developing guidelines might elect not to make a recommendation for clinical practice and to make a specific recommendation regarding the research that is needed to reduce uncertainty and clarify the trade-offs.
This example is typical of the value judgments that underlie recommendations about screening, but the same issues arise in making recommendations about treatment for both acute and chronic conditions, where it is always necessary to balance the expected benefits against the expected harms in light of the relative values attached to each important outcome, and uncertainty.

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)	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
Fluoxetine (38)											
Fluvoxamine (25)											
Nefazodone (2)											
Paroxetine (18)											
Sertraline (4)											
Venlafaxine (4)											
Transient side effects resulting in discontinuation of treatment											
Citalopram (8)	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
Fluoxetine (50)											
Fluvoxamine (27)											
Nefazodone (4)											
Paroxetine (23)											
Sertraline (6)											
Venlafaxine (5)											
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.

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.

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 more fully and carefully consider patients' values and preferences 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 (as illustrated in box 4). 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 rely, implicitly if not explicitly, 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 making accurate estimates of 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 quality of the evidence for differences in use of resources should be graded by using the criteria 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.[49] After discussion we agreed that there was moderate quality evidence for the relative effects of selective serotonin reuptake inhibitors and tricyclic antidepressants on depression severity 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 selective serotonin reuptake inhibitors (no difference in depression severity, fewer transient side effects, and fewer poisoning fatalities). Despite agreement that there seemed to be net benefits we concluded with a recommendation to "probably" use selective serotonin reuptake inhibitors, reflecting uncertainty because of the quality of the evidence. We did not have evidence of the costs of using selective serotonin reuptake inhibitors compared with tricyclics for this exercise. Had we considered costs this recommendation might have changed.

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

Conclusions

In any system that might be used to grade the quality of evidence and strength of recommendations there is a need to balance simplicity and clarity. Reducing the complexity of a system is also likely to reduce clarity, since judgments are more likely to be made implicitly rather than explicitly in simple systems. On the other hand, efforts to improve clarity and make judgments more transparent are likely to result in more complexity. In the system described here we have attempted to find a balance between simplicity and clarity. Regardless of how simple or complex a system is, judgments are always required. The approach that we have described 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.

Members of the Grades of Recommendation Assessment, Development and Evaluation (GRADE) Working Group who have contributed to this article include David Atkins, Dana Best, Peter A Briss, Martin Eccles, Yngve Falck-Ytter, Signe Flottorp, Gordon H Guyatt, Robin T Harbour, Margaret C Haugh, David Henry, Suzanne Hill, Roman Jaeschke, Gillian Leng, Alessandro Liberati, Nicola Magrini, James Mason, Philippa Middleton, Jacek Mrukowicz, Dianne O'Connell, Andrew D Oxman, Bob Phillips, Holger J Schünemann, Tessa Tan-Torres Edejer, Helena Varonen, Gunn E Vist, John W Williams Jr, Stephanie Zaza.

The National Institute for Clinical Excellence (NICE) for England and Wales and the Polish Institute for Evidence-Based Medicine (PIEBM) have provided support for meetings of the GRADE Working Group. The institutions with which members of the Working Group are affiliated have provided intramural support. Alessandro Liberati's participation in GRADE activities was supported by a grant from the Ministero Università e Ricerca Scientifica (M.I.U.R., Progetto COFIN 2001).

Contributors: All of the members of the GRADE Working Group listed above have contributed to the preparation of this manuscript and the development of the ideas contained in it, participated in at least one meeting, and read and commented on drafts of this article. GHG and ADO led the process. GEV has had primary responsibility for preparing the evidence profiles used in the pilot study and coordinating the process.

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 Canadian Task Force on the Periodic Health Examination. The periodic health examination. *CMAJ* 1979;121:1193-254.
2 Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1986;89(suppl 2):2-3S.
3 Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Arch Intern Med* 1986:146:464-5.
4 Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1989;95:2-4S.
5 Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Antithrombotic therapy consensus conference. *Chest* 1992;102(suppl 4):305-11S.
6 US Department of Health and Human Services, Public Health Service, Agency Health Care Policy and Research. *Acute pain management: operative or medical procedures and trauma.* Rockville, MD: Agency for Health Care Policy and Research Publications, 1992. (AHCPR Pub 92-0038.)
7 Gyorkos TW, Tannenbaum TN, Abrahamowicz M, Oxman AD, Scott EA, Millson ME, et al. An approach to the development of practice guidelines for community health interventions. *Can J Public Health* 1994;85(suppl 1):S8-13.
8 Hadorn DC, Baker D. Development of the AHCPR-sponsored heart failure guideline: methodologic and procedural issues. *Jt Comm J Qual Improv* 1994;20:539-54.
9 Cook DJ, Guyatt GH, Laupacis A, Sackett DL, Goldberg RJ. Clinical recommendations using levels of evidence for antithrombotic agents. *Chest* 1995;108(suppl 4):227-30S.
10 Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ, et al. Users' guide to the medical literature IX: a method for grading health care recommendations. *JAMA* 1995;274:1800-4.
11 Scottish Intercollegiate Guidelines Network (SIGN). Forming guideline recommendations. In: *A guideline developers' handbook.* Edinburgh: SIGN, 2001. (Publication No 50.) www.sign.ac.uk/guidelines/fulltext/50/section6.html (accessed 8 Feb 2004).
12 US Preventive Services Task Force. *Guide to clinical preventive services.* 2nd ed. Baltimore: Williams and Wilkins, 1996:xxxix-lv.
13 Eccles M, Clapp Z, Grimshaw J, Adams PC, Higgins B, Purves I, et al. North of England evidence based guidelines development project: methods of guideline development. *BMJ* 1996;312:760-2.
14 Centro per la Valutazione della Efficacia della Assistenza Sanitaria (CeVEAS).Schema di grading CeVEAS. http://web1.satcom.it/interage/ceveas/html/doc/45/GLICO.pdf (accessed 18 May 2004).
15 Guyatt G, Schünemann H, Cook D, Jaeschke R, Pauker S, Bucher H. Grades of recommendation for antithrombotic agents. *Chest* 2001;119:3S-7S. www.chestjournal.org/content/vol119/1_suppl/ (accessed 8 Feb 2004).
16 Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, Dawes M. Levels of evidence and grades of recommendations. Oxford: Oxford Centre for Evidence-Based Medicine. www.cebm.net/levels_of_evidence.asp (accessed 8 Feb 2004).
17 National Health and Medical Research Council. How to use the evidence: assessment and application of scientific evidence. Canberra: AusInfo, 2000. www.health.gov.au/nhmrc/publications/pdf/cp69.pdf (accessed 8 Feb 2004).
18 Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. *BMJ* 2001;323:334-6.
19 Roman SH, Silberzweig SB, Siu AL. Grading the evidence for diabetes performance measures. *Eff Clin Pract* 2000;3:85-91.
20 Woloshin S. Arguing about grades. *Eff Clin Pract* 2000;3:94-5.
21 Guyatt GH, Schünemann H, Cook D, Pauker S, Sinclair J, Bucher H, et al. Grades of recommendation for antithrombotic agents. *Chest* 2001;119: 3-7S.
22 Atkins D, Best D, Shapiro EN, eds. Third US Preventive Services Task Force: background, methods and first recommendations. *Am J Prev Med* 2001;20:3(suppl):1-108.
23 Woolf SH, Atkins D. The evolving role of prevention in health care: contributions of the US Preventive Services Task Force. *Am J Prev Med* 2001;20:3(suppl):13-20.
24 Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the process. *Am J Prev Med* 2001;20:3 (suppl):21-35.
25 Briss PA, Zaza S, Pappaioanou M, Fielding J, Wright-De Aguero L, et al. Developing an evidence-based guide to community preventive services—methods. *Am J Prev Med* 2000;18 (suppl 1):35-43.
26 Zaza S, Wright-De A, Briss PA, Truman BI, Hopkins DP, Hennessy MH, et al. Data collection instrument and procedure for systematic reviews in the guide to community preventive services. *Am J Prev Med* 2000;18(suppl 1):44-74.
27 Greer N, Mosser G, Logan G, Halaas GW. A practical approach to evidence grading. *Jt Comm J Qual Improv* 2000;26:700-12.
28 West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. *Systems to rate the strength of scientific evidence.* Rockville, MD: Agency for Healthcare Research and Quality, 2002:64-88. (AHRQ publication No 02-E016.)
29 Guyatt G, Drummond R, eds. *Users' guide to the medical literature.* Chicago, IL: AMA Press, 2002:55-154.
30 Clarke M, Oxman AD, eds. Assessment of study quality. Cochrane reviewers' handbook 4.1.5 section 6. In: *Cochrane Library.* Issue 4. Oxford: Update Software, 2002.
31 Jüni P, Altman DG, Egger M. Assessing the quality of randomised controlled trials. In: Egger M, Davey Smith G, Altman DG, eds. *Systematic reviews in health care: meta-analysis in context.* London: BMJ Books, 2001:87-121.
32 West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. *Systems to rate the strength of scientific evidence.* Rockville, MD: Agency for Healthcare Research and Quality, 2002:51-63. (AHRQ publication No 02-E016.)
33 Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials (Cochrane methodology review). In: *Cochrane Library* Issue 4. Oxford: Update Software, 2002.
34 Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;286:821-30.
35 Kleijnen J, Gøtzsche P, Kunz RA, Oxman AD, Chalmers I. So what's so special about randomisation? In: Chalmers I, Maynard A, eds. *Non-random reflections on health care research: on the 25th anniversary of Archie Cochrane's effectiveness and efficiency.* London: BMJ, 1997:93-106.
36 Lacchetti C, Guyatt G. Surprising results of randomized controlled trials. In: Guyatt G, Drummond R, eds. *Users' guide to the medical literature.* Chicago, IL: AMA Press, 2002:247-65.
37 Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA* 1998;280:605-13.

38 Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. Principal results from the women's health initiative randomized controlled trial. *JAMA* 2002;288:321-33.
39 Dans A, McAlister F, Dans L, Richardson WS, Straus S, Guyatt G. Applying results in individual patients. In: Guyatt G, Drummond R, eds. *Users' guide to the medical literature.* Chicago, IL: AMA Press, 2002:369-84.
40 McAlister F, Laupacis A, Wells G. Drug class effects. In: Guyatt G, Drummond R, eds. *Users' guide to the medical literature.* Chicago, IL: AMA Press, 2002:415-31.
41 Echt DS, Liebson PR, Mitchell LB, Peters RW, Obias-Manno D, Barker AH, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The cardiac arrhythmia suppression trial. *N Engl J Med* 1991;324:781-8.
42 Riggs BL, Hodgson SF, O'Fallon WM, Chao EY, Wahner HW, Muhs JM, et al. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. *N Engl J Med* 1990;322:802-9.
43 Worster A, Preyra I, Weaver B, Haines T. The accuracy of noncontrast helical computed tomography versus intravenous pyelography in the diagnosis of suspected acute urolithiasis: a meta-analysis. *Ann Emerg Med* 2002;40:280-6.
44 Worster A, Haines T. Does replacing intravenous pyelography with noncontrast helical computed tomography benefit patients with suspected acute urolithiasis? *Can Assoc Radiol J* 2002;53:241.
45 Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: evidence from published meta-analyses. *BMJ* 2003;326:472.
46 Bross IDJ. Pertinency of an extraneous variable. *J Chron Dis* 1967;20: 487-95.
47 Thompson DC, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. *Cochrane Database Syst Rev* 2000;(2):CD001855.
48 Devereaux PJ, Choi PT, Lacchetti C, Weaver B, Schünemann HJ, Haines T, et al. A systematic review and meta-analysis of studies comparing mortality rates of private for-profit and private not-for-profit hospitals. *CMAJ* 2002;166:1399-406.
49 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)

bmj.com 2004;328:1490

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