ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 395–400

Journal of Clinical Epidemiology

GRADE guidelines: 2. Framing the question and deciding on important outcomes

Gordon H. Guyatt[a,*], Andrew D. Oxman[b], Regina Kunz[c], David Atkins[d], Jan Brozek[a], Gunn Vist[b], Philip Alderson[e], Paul Glasziou[f], Yngve Falck-Ytter[g], Holger J. Schünemann[a]

[a]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada*
[b]*Norwegian Knowledge Centre for the Health Services, Oslo, Norway*
[c]*Academy of Swiss Insurance Medicine, University Hospital Basel, Basel, Switzerland*
[d]*Department of Veterans Affairs, Office of Research and Development, Washington, DC, USA*
[e]*Centre for Clinical Practice, National Institute for Health and Clinical Excellence (NICE), Manchester, United Kingdom*
[f]*Bond University, Gold Coast, Australia*
[g]*Division of Gastroenterology, Case and VA Medical Center, Case Western Reserve University, Cleveland, OH, USA*

Accepted 23 September 2010

Abstract

GRADE requires a clear specification of the relevant setting, population, intervention, and comparator. It also requires specification of all important outcomes—whether evidence from research studies is, or is not, available. For a particular management question, the population, intervention, and outcome should be sufficiently similar across studies that a similar magnitude of effect is plausible. Guideline developers should specify the relative importance of the outcomes before gathering the evidence and again when evidence summaries are complete. In considering the importance of a surrogate outcome, authors should rate the importance of the patient-important outcome for which the surrogate is a substitute and subsequently rate down the quality of evidence for indirectness of outcome.

Keywords: GRADE; PICO; Patient-important outcomes; Surrogate; Guideline development; Quality of evidence; Indirectness

1. Introduction

In the first article of this series, we introduced GRADE and the GRADE evidence profile and summary-of-findings tables that facilitate clinical decisions. This second article discusses GRADE's approach in framing the relevant questions for systematic reviews and guidelines, choosing the relevant outcomes and deciding on their relative importance. We focus on conceptual issues: later articles will address who exactly should take on what roles.

2. Structured questions of patient management

This article will focus on questions about the effects of interventions. Guideline developers will, however, usually have important questions about prognosis, prevalence, and other types of questions that require a different framing structure than management issues (Box 1).

3. Framing questions involves specifying patients, interventions, comparators, and outcomes, and sometimes setting

A well-accepted methodology associated with framing of questions addressing alternative management strategies in systematic reviews mandates carefully specifying the patient population, the intervention of interest, the comparator, and the outcomes of interest. The value of the methodology—popularly known as PICO (patient/intervention/comparator/outcome)—in helping achieve focused recommendations is increasingly recognized not only by systematic review authors but also by guideline developers [1].

A guideline question often involves another specification: the setting in which the guideline will be implemented. For instance, guidelines intended for resource-rich environments will often be inapplicable to resource-poor environments. In the first article in this series, we presented an evidence profile describing the impact of antibiotics on otitis media. The

The GRADE system has been developed by the GRADE Working Group. The named authors drafted and revised this article. A complete list of contributors to this series can be found on the JCE Web site at www.jclinepi.com.

* Corresponding author. Department of Clinical Epidemiology and Biostatistics, CLARITY Research Group, Room 2C12, 1200 Main Street, West Hamilton, Ontario L8N 3Z5, Canada. Tel.: +905-524-9140; fax: +905-524-3841.

E-mail address: guyatt@mcmaster.ca (G.H. Guyatt).

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2010.09.012

Key points

GRADE requires a clear specification of the relevant setting, population, intervention, comparator(s), and outcomes.

Outcomes of interest should be those important to patients: if patient-important outcomes are represented by a surrogate, they will frequently require rating down the quality of evidence for indirectness.

Questions must be sufficiently specific: across the range of populations, interventions, and outcomes, a more or less similar effect must be plausible.

For a guideline, an initial rating of the importance of outcomes should precede the review of the evidence, and this rating should be confirmed or revised following the evidence review.

Box 1 The role of questions of prognosis in guidelines

GRADE does not provide a formal structure for evaluating the quality of evidence underlying questions of prognosis. Nevertheless, they are often important for guideline development. For example, addressing interventions that may influence the outcome of influenza or multiple sclerosis will require establishing the natural history of the conditions. This will involve specifying the population (influenza or new-onset multiple sclerosis) and the outcome (mortality or relapse rate and progression). Such questions of prognosis may be refined to include multiple predictors, such as age, gender, or severity. The answers to these questions will be an important background for formulating recommendations and interpreting the evidence about the effects of treatments. In particular, guideline developers need to decide whether the prognosis of patients in the community is similar to those studied in the trials and whether there are important prognostic subgroups that they should consider in making recommendations.

results apply to high- and middle-income countries, in which the risk of progression to mastoiditis is very low.

The most challenging decision in framing the question is how broadly the patients and intervention should be defined. For example, in addressing the effects of antiplatelet agents on vascular disease, one might include only patients with transient ischemic attacks; those with ischemic attacks and strokes; or those with any vascular disease (cerebro-, cardio-, or peripheral vascular disease). The intervention might be a relatively narrow range of doses of aspirin, all doses of aspirin, or all antiplatelet agents.

On what basis should systematic-review authors or guideline developers make this decision? The underlying biology must suggest that, across the range of patients and interventions, it is plausible that the magnitude of effect on the key outcomes is more or less the same. If that is not the case, the review or guideline will generate misleading estimates for at least some subpopulations of patients and interventions.

For instance, if antiplatelet agents differ in effectiveness in those with peripheral vascular disease vs. those with myocardial infarction (as one study of clopidogrel vs. aspirin that enrolled patients from both populations suggested [2]), a single estimate across the range of patients and interventions will not well serve the decision-making needs of patients and clinicians. The same will be true if different antiplatelet agents have differing magnitudes of effect.

Often, and appropriately, systematic reviews deal with the potentially vexing question of what breadth of population or intervention to choose by starting with a broad question but including a priori specification of subgroup effects that may explain any heterogeneity they find. These hypotheses may apply to patients (e.g., effects differ in those with transient ischemic attacks and strokes vs. those with coronary or peripheral vascular diseases) or interventions (e.g., high vs. low doses of aspirin or aspirin vs. other antiplatelet agents). A priori hypotheses may also relate to the choice of comparator (e.g., effects of amiodarone on conversion to sinus rhythm in patients with atrial fibrillation differ depending on whether the comparator is placebo or an active agent unlikely to influence return to sinus rhythm [3]); the outcome (e.g., the effect of an antihypertensive agent differs on vascular events in the cerebral or myocardial circulation); or methodology (e.g., high-quality studies yield different effects than low-quality studies). We deal with the issue of subgroup effects in much more detail in a subsequent article in this series [4].

Sometimes, there are multiple comparators to an intervention, and this raises particular challenges. For example, the European Society of Cardiology makes recommendations for use of anticoagulants in patients with non-ST elevation acute coronary syndromes receiving conservative (noninvasive) management [5]. Fondaparinux receives a 1A, heparin a 1C, and enoxaparin a 2A/B. Presumably, these are recommendations for use of these agents vs. not using any anticoagulants. But do they also imply a gradient of preference of fondaparinux over heparin over enoxaparin?

Clarity in choice of the comparator makes for interpretable guidelines—and lack of clarity can cause confusion. Sometimes, the comparator is obvious—when, however, it is not, guideline panels should specify the comparator explicitly. In particular, when multiple agents are involved, they should specify whether the recommendation is suggesting that all agents are equally recommended or that some agents are recommended over others.

4. Ensuring the question framing is appropriately specific

Because the relative risk associated with an intervention vs. a specific comparator is usually similar across a wide variety of baseline risks, it is usually appropriate for systematic reviews to generate single pooled estimates of relative effects across a wide range of patient subgroups [6,7,8]. For instance, the relative risk reduction in vascular events associated with statins is very similar in those with and without underlying vascular disease; the relative risk reduction associated with warfarin vs. both no-antithrombotic therapy and aspirin appears similar across patients with atrial fibrillation at low and higher risk of stroke.

Recommendations, however, may differ across subgroups of patients at different baseline risk of an outcome, despite there being a single relative risk that applies to all of them. For instance, the case for warfarin therapy—associated with both inconvenience and a higher risk of serious bleeding—is much stronger in atrial fibrillation patients at substantial vs. minimal risk of stroke [9]. Absolute risk reductions are greater in higher-risk patients, warranting taking a higher risk of side effects and enduring inconvenience. Evidence quality may also differ across subgroups, and this may mandate differing recommendations (higher likelihood of recommending an intervention, or making a stronger recommendation, when evidence is of higher quality). Thus, guideline panels must often define separate questions (and produce separate evidence summaries) for high- and low-risk patients, and patients in whom quality of evidence differs, included in a single meta-analysis.

5. Specification of outcomes: ensuring comprehensiveness

Many, if not most, systematic reviews fail to address some key outcomes, particularly harms, associated with an intervention. Systematic reviews may even focus on a single outcome (e.g., the impact of statins on stroke [10] or vitamin D on nonvertebral fractures [11]).

Guideline panels do not have this luxury. Sensible recommendations require consideration of all outcomes that are important to patients. In addition, they may require consideration of outcomes that are important to others, including the use of resources paid for by third parties; impacts on those who care for patients; and public health impacts (e.g., the spread of infections or antibiotic resistance).

If evidence is lacking for an important outcome, this should be acknowledged, rather than ignoring the outcome—that uncertainty may have a bearing on the ultimate recommendation. Deciding on recommendations regarding statins for patients at risk of stroke involves considering effects not only on stroke but on other vascular events as well as adverse effects of rhabdomyolysis and liver injury; recommendations regarding vitamin D must consider both vertebral fractures and putative benefits in cancer prevention. Outcomes that panels need to consider for most recommendations will include morbid and mortal events and adverse effects. Often, other outcomes, such as hospitalization, function, disability, quality of life, inconvenience, and resource use, will also be important.

Because most systematic reviews do not summarize the evidence for all important outcomes, guideline panels must often either use multiple systematic reviews from different sources or conduct their own systematic reviews.

6. Outcome importance: three categories

Guideline panels using GRADE will consider the importance of outcomes in three steps (Table 1). We will address the first two steps in this article. In subsequent articles, we will address the third step—making judgments about the balance between the desirable and undesirable effects of an intervention.

Guideline developers must, and authors of systematic reviews ideally will, specify all potential patient-important outcomes as the first step in their endeavor. Those using GRADE for guideline development will also make a preliminary classification of outcomes into those that are critical, those that are important but not critical, and those of limited importance. The first two classes of evidence will bear on guideline recommendations; the third may or may not. Guideline developers may choose to rate outcomes numerically on a 1−9 scale (7−9, critical; 4−6, important; and 1−3, of limited importance) to distinguish between importance categories (Fig. 1). Ranking outcomes by their relative importances can help to focus attention on those outcomes that are considered most important and help to resolve or clarify disagreements. For instance, Fig. 1 suggests that flatulence is of little importance to patients. If flatulence is persistent or severe, this may not be the case.

Later in this series, we will elaborate on the need to distinguish between critical and important-but-not-critical outcomes. For now, it would suffice to say that decisions regarding the overall quality of evidence supporting a recommendation may depend on which outcomes are designated as critical for making the decision (e.g., those rated 7, 8, or 9, on the 9-point scale mentioned earlier) and which are not.

For instance, a guideline panel decides that high-quality evidence supports all outcomes but one, and that only low-quality evidence is available for the remaining outcome. If that remaining outcome is critical, the overall quality of evidence will be designated as low quality. If the panel feels that the remaining outcome is important but not critical, the overall rating of quality of evidence for the associated recommendation will be of high quality.

7. Outcome importance: influence of perspective

Importance of outcomes is likely to vary within and across cultures or when considered from the perspective of patients, clinicians, or policy makers. Guideline panels

Table 1
Three steps for considering the relative importance of outcomes

Step	What	Why	How	Evidence
1	Preliminary classification of outcomes as critical, important but not critical, or low importance, before reviewing the evidence	To focus attention on those outcomes that are considered most important when searching for and summarizing the evidence and to resolve or clarify disagreements	By asking panel members and possibly patients or members of the public to identify important outcomes, judging the relative importance of the outcomes and discussing disagreements. Conducting a systematic review of the relevant literature	These judgments can draw on the experience of the panel members, patients, and members of the public. Prior knowledge of the research evidence or, ideally, a systematic review of that evidence is likely to be helpful
2	Reassessment of the relative importance of outcomes after reviewing the evidence	To ensure that important outcomes identified by reviews of the evidence that were not initially considered are included and to reconsider the relative importance of outcomes in light of the available evidence	By asking the panel members (and, if relevant, patients and members of the public) to reconsider the relative importance of the outcomes included in the first step and any additional outcomes identified by reviews of the evidence	Experience of the panel members and other informants and systematic reviews of the effects of the intervention
3	Judging the balance between the desirable and undesirable effects of an intervention	To make a recommendation and to determine the strength of the recommendation	By asking the panel members (and, if relevant, patients and members of the public) to judge the balance between the desirable and undesirable effects using a balance sheet (summary-of-findings table) and, if relevant, using a decision analysis or an economic analysis	Experience of the panel members and other informants, systematic reviews of the effects of the intervention, evidence of the value that patients attach to key outcomes (if relevant and available), and decision analyses or economic analyses (if relevant and available)

must decide what perspective they are taking. Although different panels may elect to take different perspectives (e.g., that of individual patients, that of a third-party payer, or a societal perspective), the relative importance given to outcomes should reflect the perspective of those who are affected. When the target audiences for a guideline are clinicians and the patients they treat, the perspective would generally be that of the patient. A subsequent article in this series will address the issue of perspective from the point of view of resource use.

8. Importance of outcomes: using evidence

At the time of writing, a guideline panel sponsored by the American College of Chest Physicians (ACCP) is developing the ninth iteration of the ACCP antithrombotic guidelines. As part of this process, the group has conducted a systematic review of the evidence relating to patients' values and preferences for antithrombotic therapy. Insights from this review have included the considerable variability of patients' values, the limited burden of warfarin therapy that most patients experience, and the relative weighting of stroke and serious bleeding outcomes.

In the absence of such evidence, clinicians can use their prior interactions with patients to make deductions about patient values and preferences. For instance, in the eighth iteration of the antithrombotic guidelines, the panelists responsible for the pregnancy chapter wrote ''anecdotal evidence suggests that many, though not all women, give higher priority to the impact of any treatment on the health of their unborn baby than to effects on themselves.''

9. Outcome importance: missing evidence and surrogate outcomes

Systematic reviews—though they may reflect on the implications of what is measured and what is not measured—are limited to preparing quantitative summaries of outcomes that the investigators have included in their studies. Not infrequently, outcomes of most importance to patients remain unexplored. For example, in type 2 diabetes, clinical trials have failed to adequately address the long-term impact of alternative management strategies on diabetic complications of micro- and macrovascular disease, and neuropathic complications [12]; this omission is unlikely to be corrected in the near future [13].

When important outcomes are relatively infrequent, or occur over long periods of time, clinical trialists often choose to measure substitutes, or surrogates, for those outcomes. It may be tempting—though we would argue misguided—for guideline developers to assume that intervention impact on surrogates reflects impact on patient-important outcomes. Because of the many instances in which this assumption has proven wrong [14], guideline developers using GRADE

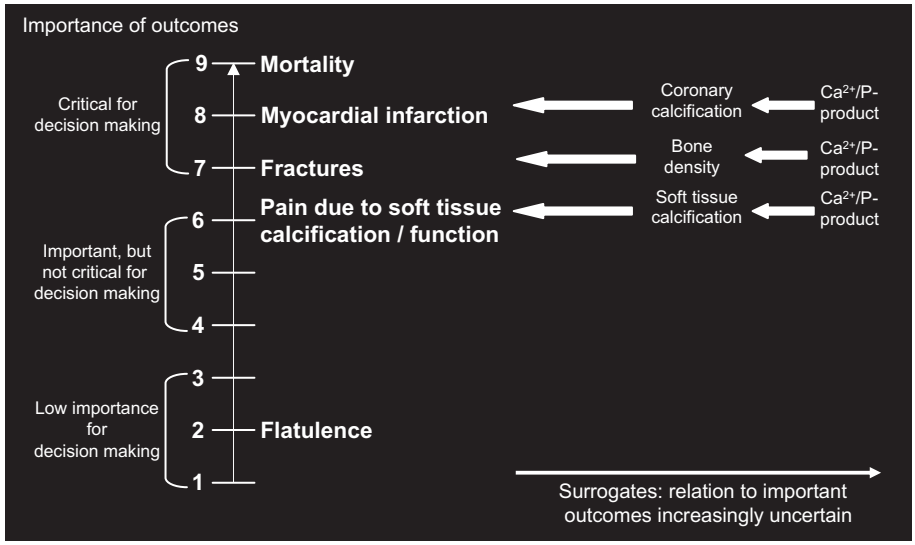

Fig. 1. Hierarchy of outcomes according to their importance to assess the effect of phosphate-lowering drugs in patients with renal failure and hyperphosphatemia.

will specify patient-important outcomes and, if necessary, the surrogates they are using to substitute for those important outcomes.

Consider, for instance, a guideline panel addressing the use of novel agents to lower phosphate in patients with renal failure and hyperphosphatemia. What are the intended effects of the intervention? The obvious answer may be to lower serum phosphate, but the more appropriate answer is to reduce mortality, myocardial infarction, fractures, and pain because of soft-tissue calcification (Fig. 1). Trials may, however, measure only surrogates related specifically to each of these outcomes (coronary calcification, bone density, or radiological manifestation of tissue calcification) or even more distant, generic surrogates (calcium phosphate product) (Fig. 1).

Guideline developers should consider surrogate outcomes only when high-quality evidence regarding important outcomes is lacking. When such evidence is lacking, guideline developers may be tempted to list the surrogates as their measures of outcome. This is not the approach GRADE recommends. Rather, they should specify the important outcomes and the associated surrogates they must use as substitutes. As we will describe later in this series, the necessity to substitute the surrogate may ultimately lead to rating down the quality of the evidence because of indirectness.

10. Outcome importance: preliminary and definitive ratings

Although it is worthwhile to specify critical and important outcomes before beginning the review of the evidence, results of that review may influence judgments about the importance of the outcomes. We describe two situations in which results of the evidence review may modify the selection of relevant outcomes or their relative importance as follows.

1. A potential benefit on a particular outcome, initially judged critical, may no longer be critical on review of the results. This will be the case if, given other established benefits, we would still be enthusiastic about the intervention in the absence of a demonstrated benefit on the outcome in question.

   Consider, for instance, a screening intervention, such as screening for aortic abdominal aneurysm. Initially, a guideline panel is likely to consider the intervention's impact on all-cause mortality as critical. Let us say, however, that the evidence summary establishes an important reduction in cause-specific mortality from abdominal aortic aneurysm but fails to definitively establish a reduction in all-cause mortality. The reduction in cause-specific mortality may be judged sufficiently compelling that, even in the absence of a demonstrated reduction in all-cause mortality (which may be undetected because of random error from other causes of death), the screening intervention is clearly worthwhile. All-cause mortality then becomes less relevant and ceases to be a critical outcome.

   This reasoning requires careful consideration of two potential problems. First, we must be reasonably certain that there is no increase in all-cause mortality associated with the intervention (as is highly likely with ultrasound screening for aneurysms). Second, the magnitude of the absolute benefit on disease-specific mortality must be sufficiently large that the net benefit of the intervention is clear without a demonstrated reduction in all-cause mortality. Guideline authors should, in general, note the reasoning underlying the designation of critical and important outcomes and, in particular, judgments, such as those described earlier.
2. Any new intervention may be associated with adverse effects that are not initially apparent. Indeed, over

a quarter of a century, important unexpected toxicity has emerged in approximately 20% of the U.S. Food and Drug Administration−approved drugs [15]. Thus, one might consider ''as-yet-undiscovered toxicity'' as an important adverse consequence of any new drug.

Such toxicity becomes critical only when sufficient evidence of its existence emerges. For instance, myocardial infarction might, when the drugs were initially marketed, have been one among a long list of speculative adverse effects (e.g., autoimmune syndromes, bone marrow suppression, renal failure), possibly associated with the use of COX-2 inhibitors. When evidence of increased rate of myocardial infarction with COX-2 inhibitors emerged, it then became a critical outcome.

The tricky part of this judgment is how frequently the adverse event must occur and how plausible the association with the intervention must be before it becomes a critical outcome. For instance, an observational study found a previously unsuspected association between sulfonylurea use and cancer-related mortality [16]. Should cancer deaths now be an important, or even a critical, endpoint when considering sulfonylurea use in patients with type 2 diabetes?

As is repeatedly the case, we cannot offer hard-and-fast rules for these judgments. What GRADE does is label the issues involved and permit a transparent and explicit accounting of the judgments involved. Guideline panel members can then debate the issues, and guideline users make their own assessment of the appropriateness of the panel's conclusions.

References

[1] Schunemann HJ, Cook D, Guyatt G. Methodology for antithrombotic and thrombolytic therapy guideline development: American College of Chest Physicians Evidence-based Clinical Practice Guidelines. (8th edition). Chest 2008;133(6 Suppl):113S−22S.

[2] CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). Lancet 1997;348:1329−39.

[3] Letelier LM, Udol K, Ena J, Weaver B, Guyatt GH. Effectiveness of amiodarone for conversion of atrial fibrillation to sinus rhythm: a meta-analysis. Arch Intern Med 2003;163:777−85.

[4] Guyatt G, et al. Grade guidelines: 7. Rating the quality of evidence: inconsistency. J Clin Epidemiol 2010. In press.

[5] Eikelboom J, Guyatt G, Hirsh J. Guidelines for anticoagulant use in acute coronary syndromes. Lancet 2008;371:1559−61.

[6] Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the ''number needed to treat''? An empirical study of summary effect measures in meta-analyses. Int J Epidemiol 2002;31:72−6.

[7] Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Stat Med 1998;17:1923−42.

[8] Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002;21:1575−600.

[9] Singer DE, Albers GW, Dalen JE, Fang MC, Go AS, Halperin JL, et al. Antithrombotic therapy in atrial fibrillation: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. (8th edition). Chest 2008;133(6 Suppl):546S−92S.

[10] Bucher HC, Griffith LE, Guyatt GH. Effect of HMGcoA reductase inhibitors on stroke. A meta-analysis of randomized, controlled trials. Ann Intern Med 1998;128:89−95.

[11] Bischoff-Ferrari HA, Willett WC, Wong JB, Giovannucci E, Dietrich T, Dawson-Hughes B. Fracture prevention with vitamin D supplementation: a meta-analysis of randomized controlled trials. JAMA 2005;293:2257−64.

[12] Montori VM, Gandhi GY, Guyatt GH. Patient-important outcomes in diabetes—time for consensus. Lancet 2007;370:1104−6.

[13] Gandhi GY, Murad MH, Fujiyoshi A, Mullan RJ, Flynn DN, Elamin MB, et al. Patient-important outcomes in registered diabetes trials. JAMA 2008;299:2543−9.

[14] Bucher H, et al. Surrogate outcomes. In: Guyatt G, et al, editors. The users' guides to the medical literature: a manual for evidence-based clinical practice. New York, NY: McGraw-Hill; 2008.

[15] Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SM, Bor DH. Timing of new black box warnings and withdrawals for prescription medications. JAMA 2002;287:2215−20.

[16] Bowker SL, Majumdar SR, Veugelers P, Johnson JA. Increased cancer-related mortality for patients with type 2 diabetes who use sulfonylurea or insulin. Diabetes Care 2006;29:254−8.