ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 726−735

Journal of Clinical Epidemiology

GRADE guidelines: 15. Going from evidence to recommendation—determinants of a recommendation's direction and strength

Jeffrey C. Andrews[a,*], Holger J. Schünemann[b,c], Andrew D. Oxman[d], Kevin Pottie[e], Joerg J. Meerpohl[f,g], Pablo Alonso Coello[h,i], David Rind[j], Victor M. Montori[k], Juan Pablo Brito[k], Susan Norris[l], Mahmoud Elbarbary[m], Piet Post[n], Mona Nasser[o], Vijay Shukla[p], Roman Jaeschke[c], Jan Brozek[b], Ben Djulbegovic[q,r], Gordon Guyatt[b,c]

[a]*Vanderbilt Evidence-based Practice Center, Vanderbilt University, #27166-719 Thompson Lane, Nashville, TN 37204-3195, USA*
[b]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[c]*Department of Medicine, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[d]*Norwegian Knowledge Centre for the Health Services, PO Box 7004, St. Olavs plass, Oslo 0130, Norway*
[e]*Department of Family Medicine, University of Ottawa, Ottawa, Canada*
[f]*German Cochrane Center, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Berliner Allee 29, 79110 Freiburg, Germany*
[g]*Division of Pediatric Hematology and Oncology, Center for Pediatrics and Adolescent Medicine, University Medical Center Freiburg, Mathildenstrasse 1, 79106 Freiburg, Germany*
[h]*Iberoamerican Cochrane Center, CIBER de Epidemiología y Salud Pública, IIB, Sant Pau, Barcelona 08041, Spain*
[i]*Epidemiology and Public Health CIBER (CIBERESP), Hospital de la Sant Pau, Creu i Sant Pau, Barcelona 08041, Spain*
[j]*Harvard Medical School, Beth Israel Deaconess Medical Center, Healthcare Associates-E/Shapiro 6, 330 Brookline Aveen, Boston, MA 02215, USA*
[k]*Mayo Clinic, 200 1st SW St, Rochester, MN 55905, USA*
[l]*Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239-3098, USA*
[m]*King Saud University for Health Sciences, Riyadh, Saudi Arabia*
[n]*Post Voor Zorg, Delft, The Netherlands*
[o]*Peninsula College of Medicine and Dentistry, Universities of Exeter and Plymouth, The John Bull Building, Tamar Science Park, Plymouth, PL68BU, UK*
[p]*Canadian Agency for Drugs and technologies in Health (CADTH), 600-865 Carling Avenue, Ottawa, Ontario K1S 5S8, Canada*
[q]*Division and Center for Evidence-Based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL, USA*
[r]*H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA*

Accepted 3 February 2013; Published online 6 April 2013

Abstract

In the GRADE approach, the strength of a recommendation reflects the extent to which we can be confident that the composite desirable effects of a management strategy outweigh the composite undesirable effects.

This article addresses GRADE's approach to determining the direction and strength of a recommendation. The GRADE describes the balance of desirable and undesirable outcomes of interest among alternative management strategies depending on four domains, namely estimates of effect for desirable and undesirable outcomes of interest, confidence in the estimates of effect, estimates of values and preferences, and resource use. Ultimately, guideline panels must use judgment in integrating these factors to make a strong or weak recommendation for or against an intervention. © 2013 Elsevier Inc. All rights reserved.

Keywords: GRADE; Quality of evidence; Strength of evidence; Guideline development; Recommendation; Evidence

1. Introduction

In prior articles in this series devoted to the GRADE approach to systematic reviews and practice guidelines, we have dealt with the process before developing recommendations, namely framing the question and choosing critical and important outcomes [1], rating the confidence in effect estimates for each outcome [2−8], dealing with resource

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 *Journal of Clinical Epidemiology* web site.

* Corresponding author. Tel.: (615) 343-5700.
E-mail address: jeff.andrews@vanderbilt.edu (J.C. Andrews).

0895-4356/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jclinepi.2013.02.003

use [9], rating the confidence in effect estimates across outcomes [10], and creating an evidence profile and a Summary of Findings table [11−13]. The immediately previous article described GRADE's approach to classifying the strength and direction of recommendations and discussed the implications of strong and weak recommendations, and the options for presentation and wording [14]. The present article presents GRADE's approach to moving from evidence to recommendations. As we did in the previous article, we will refer to guideline developers as ''the panel.''

1.1. Globalizing evidence and localizing decisions

The pithy summary by Eisenberg [15] on the relationship between evidence and recommendations, ''globalize the evidence, localize the decisions,'' provides fundamental guidance for those working to produce evidence-based recommendations [15]. Summaries of evidence regarding alternative management strategies from the medical literature should ideally be very similar, no matter the site of the application of the recommendation.

Rating of confidence in estimates of effect (quality of evidence) may, however, differ for a variety of reasons. First, desirable and undesirable outcomes may be valued differently, leading to different thresholds of acceptability. This could lead to different judgments regarding imprecision, as we have highlighted in the article in this series dealing with imprecision [5].

Second, differences in values and preferences could lead to differences in the overall balance of desirable and undesirable outcomes and the rating of confidence in estimates: an outcome judged as critical by one panel (and thus included in the rating of overall confidence in estimates) may be judged important but not critical by another (and thus not included in the overall rating).

Finally, ratings of confidence may also differ as a result of uncertainties in the risk profile of untreated populations (baseline risk). We may be very confident of baseline risk in one setting but not at all confident in another. This could lead to rating down confidence in estimates for indirectness.

Continued rapid uptake of GRADE by organizations that produce systematic summaries of evidence will greatly facilitate the production of transparent evidence summaries. If organizations work together to produce summaries, there will be an enormous gain in efficiency [16]—even if, in the end, judgments about confidence in estimates will differ across settings, for reasons described in the preceding paragraphs.

We now turn to a systematic presentation of the determinants of direction and strength of recommendations.

2. Determinants of direction and strength of recommendations

GRADE has identified six determinants of the direction and strength of recommendations, namely the magnitude of estimates of effect of the interventions on important outcomes, confidence in those estimates, estimates of typical values and preferences, confidence in those estimates, variability of values and preferences, and resource use. In the presentation here, we will present these six determinants in four domains. We package magnitude of effect and typical values and preferences together with the label balance of desirable and undesirable consequences or ''trade-offs.'' We also include uncertainty regarding typical values, and variability in values, in a single domain (Table 1).

Alternative groupings may work better, depending on the circumstances. We believe that the approach we present here is best for presenting the rationale for the recommendations to the guideline consumer audience. In developing recommendations, panels may want to keep all six determinants separate or group the three values and preferences determinants together.

Ultimately, guideline panels must integrate these six determinants to make a strong or weak recommendation for or against an intervention. Table 2 illustrates how the elements of the GRADE framework for moving from evidence to recommendations can be applied in making strong and weak recommendations, and Table 3 provides an example of the application in the management of chronic obstructive pulmonary disease.

2.1. Trade-offs between desirable and undesirable consequences of alternative management strategies

When we consider the balance between desirable and undesirable outcomes (''trade-offs''), we are considering two domains. The first is our best estimates of the magnitude of desirable effects and the undesirable effects. If a guideline panel has adhered to the GRADE process, they will find the best estimates of effect in the evidence profiles that they have prepared or accessed.

The second element that determines the balance among desirable and undesirable outcomes is the typical values that patients—or a population—apply to those outcomes. This can be otherwise conceptualized as the relative preferences for those outcomes—and thus the term we generally use, values and preferences (Box 1).

Ideally, to inform estimates of typical patient values and preferences, guideline panels will conduct or identify systematic reviews of relevant studies of patient values and preferences [18]. Given the paucity of empirical examinations of patients' values and preferences, however, well-resourced guideline panels will usually complement such studies with consultation with individual patients and patients' groups. The panel should discuss whose values these people represent, namely representative patients, a defined subset of patients, or representatives of the general population.

For example, the Canadian Collaboration for Immigrant and Refugees Health (CCIRH) guidelines sought to advance understanding of immigrant patient perspectives in

Table 1. Domains that contribute to the strength of a recommendation

Domains that contribute to the strength of a recommendation	Comment
Balance between desirable and undesirable outcomes (estimated effects), with consideration of values and preferences (estimated typical) (trade-offs)	The larger the differences between the desirable and undesirable consequences, the more likely a strong recommendation is warranted. The smaller the net benefit and the lower certainty for that benefit, the more likely a weak recommendation is warranted
Confidence in the magnitude of estimates of effect of the interventions on important outcomes (overall quality of evidence for outcomes)	The higher the quality of evidence, the more likely a strong recommendation is warranted
Confidence in values and preferences and variability	The greater the variability in values and preferences, or uncertainty in values and preferences, the more likely a weak recommendation is warranted
Resource use	The higher the costs of an intervention (the more resources consumed), the less likely a strong recommendation is warranted

two ways, namely they searched and synthesized evidence for immigrant perspectives in relation to each health condition, and worked closely with a community-based organization representing 18 ethnic groups to inform perceptions of immigrant patient perspectives [19]. Less well-resourced panels, without systematic reviews of values and preferences or consultation with patients and patient groups, must rely on unsystematic reviews of the available literature and their clinical experience of interactions with patients. How well such estimates correspond to true typical values and preferences is likely, in any particular situation, to be uncertain.

Whatever the source of estimates of typical values and preferences, explicit, transparent statements of the panel's choices are imperative. For example, in their recommendation regarding unmet contraceptive needs, the CCIRH attributed more value to supporting informed choice (empowerment) and less value to concern about causing couple and family discord [19]. Clinicians recognizing a family in which avoiding discord is paramount will therefore be aware that the recommendation is in that instance not appropriate.

Maximal explicitness requires quantification. For example, in the ninth iteration of the American College of Chest Physicians Antithrombotic Guidelines, the panel specified that they considered typical patients would value preventing one stroke equivalent to avoiding three serious gastrointestinal bleeds [18,20].

Having established their best estimates of typical values and preferences, a panel is in a position to assess the trade-off between the desirable and undesirable outcomes of an intervention vs. a comparator. The larger the gradient between the desirable and undesirable effects, the higher the likelihood that a panel will provide a strong recommendation. For example, the very large gradient between the benefits of low dose aspirin on reductions in death and

Table 2. Examples of strong and weak recommendation determinants

Factor	Example of strong recommendation	Example of weak recommendation
Balance between desirable and undesirable consequences of alternative management strategies. The closer the balance, the less likely a strong recommendation	Aspirin following myocardial infarction reduces mortality with minimal toxicity, inconvenience, and cost	Anticoagulation vs. aspirin in patients with atrial fibrillation with a $CHADS_2$ score of 1 (moderate risk of stroke); benefit in stroke reduction closely balanced with increased bleeding risk
Confidence in estimates of effect (quality of evidence). The lower the confidence, the less likely a strong recommendation	Many high quality randomized trials have shown the benefit of inhaled steroids in asthma	Only case series have examined the utility of pleurodesis in pneumothorax
Uncertainty or variability in values and preferences. The less the confidence in estimates of typical values and preferences, and the greater the variability, the less likely a strong recommendation	Relative confidence: evidence from empirical studies shows that patients place a substantially higher value on avoiding a debilitating stroke than on avoiding a serious gastrointestinal bleed	Uncertainty: there is no empirical evidence regarding the relative value patients place on avoiding a postoperative bleed that requires reoperation vs. a postoperative serious but nonfatal pulmonary embolus
	Little variability: young patients with lymphoma will invariably place a higher value on the life-prolonging effects of chemotherapy than on avoiding treatment toxicity	Greater variability: some older patients with lymphoma will place a higher value on the life-prolonging effects of chemotherapy than on avoiding treatment toxicity but others will not
Resource use. The higher the resource use, the less likely a strong recommendation	The low cost of aspirin vs. no antithrombotic prophylaxis against stroke in patients with transient ischemic attacks	The high cost of clopidogrel and of combination dipyridamole and aspirin vs. aspirin as prophylaxis against stroke in patients with transient ischemic attacks

Table 3. Evidence to recommendation framework: enhancing transparency when moving from evidence to recommendations

Question/recommendation: Should pulmonary rehabilitation vs. usual community care be used for COPD with recent exacerbation?			
Population: Patients with COPD and recent exacerbation of their disease			
Intervention: Pulmonary rehabilitation vs. no rehabilitation			
Setting (if relevant): Outpatient			
Decision domain	**Judgment**	**Reason for judgment**	**Subdomains influencing judgment**
Balance of desirable and undesirable outcomes Given the best estimate of typical values and preferences, are you confident that the benefits outweigh the harms and burden or vice versa?	Yes ☒ No ☐	The desirable consequences are substantial (including substantial reduction in hospitalization, small but important reduction in mortality, and improvement in quality of life that exceeds the minimal important difference) and valued highly. The undesirable consequences, inconvenience, and burden are relatively minor and associated with minimal disutility.	Baseline risk for desirable and undesirable outcomes: • Is the baseline risk similar across subgroups? • Should there be separate recommendations for subgroups? Relative risk for benefits and harms: • Are the relative benefits large? • Are the relative harms large? Requirement for modeling: • Is there a lot of extrapolation and modeling required for these outcomes? Typical values: • What are the typical values? • Are there differences in the relative value of the critical outcomes?
Confidence in estimates of effect (quality of evidence) Is there high or moderate quality evidence?	Yes ☒ No ☐	⊕⊕⊕O There is moderate-(mortality, function, and quality-of-life outcomes)-to-high (hospitalizations) quality evidence for the desirable consequences, and quality evidence for the undesirable (burden)	Confidence in estimates of benefits and downsides, confidence in estimates of resource use. Consider all critical outcomes, including the possibility that some may not be measured. Key reasons for rating evidence down or rating up
Values and preferences Are you confident about the typical values and preferences and are they similar across the target population?	Yes ☒ No ☐	We can be confident that patients place a high value on avoiding hospitalizations and mortality as well as improving quality of life and a low value on avoiding the inconvenience associated with rehabilitation. We can be confident that these values vary little among patients with chronic respiratory disease.	Source of typical values (panel or study of general population or patients) Source of estimates of variability and extent of variability Method for determining values satisfactory for this recommendation
Resource implications Are the resources worth the expected net benefit from following the recommendation?	Yes ☒ No ☐	There are resources required to provide pulmonary rehabilitation but these are balanced by decreased resource needs as a result of decreased hospitalizations and net cost is well worth it given the desirable outcomes.	What are the costs per resource unit? Feasibility: • Is this intervention generally available? Opportunity cost: • Is this intervention and its effects worth withdrawing or not allocating resources from other interventions Differences across settings: • Is there lots of variability in resource requirements across settings?
Overall strength of recommendation	Strong	The guideline panel recommends that patients with recent exacerbations of their COPD undergo pulmonary rehabilitation (Note: this is a hypothetical recommendation developed for this article and not intended for clinical decision making).	
Evidence to recommendation synthesis	The moderate-to-high confidence in the moderate-to-large magnitude of effects on highly valued outcomes, and the moderate-to-high confidence that undesirable outcomes are modest and their avoidance not highly valued suggest a strong recommendation.		

Abbreviation: COPD, chronic obstructive pulmonary disease.

Box 1 Terminology for "values and preferences"

Values and preferences is an overarching term that includes patients' perspectives, beliefs, expectations, and goals for health and life [17]. More precisely, they refer to the processes that individuals use in considering the potential benefits, harms, costs, limitations, and inconvenience of the management options in relation to one another. For some, the term "values" has the closest connotation to these processes. For others, the connotation of "preferences" best captures the notion of choice. Thus, we use both words together to convey the concept.

recurrent myocardial infarction (MI) after an MI [21] and the undesirable consequences of minimal side effects and costs make a strong recommendation very likely (Table 2).

In contrast, the narrower the magnitude of the gradient between desirable and undesirable consequences, the higher the likelihood that a guideline panel will make a weak recommendation. For instance, consider the choice of immunomodulating agents, namely cyclosporine and tacrolimus in kidney transplant recipients [22]. Tacrolimus results in better graft survival (a highly valued outcome), but at the important cost of a higher incidence of diabetes (the long-term complications of which can be devastating).

Table 2 presents a second example of a close trade-off in which patients with atrial fibrillation typically are more stroke averse than bleeding averse. If, however, the risk of stroke is sufficiently low, the trade-off between stroke reduction and increase in bleeding risk with anticoagulants is closely balanced.

Without considering the associated values and preferences, assessing large vs. small magnitude of effects may be misleading. For instance, in patients with cancer, chemotherapeutic agents may have large (albeit temporary) adverse effects such as nausea, fatigue, hair loss, and paresthesias. The chemotherapy may have only a small effect on reducing mortality. Despite the discrepancy in magnitude of effect, most patients may choose chemotherapy because of the very high value they place on a small mortality reduction.

2.2. *Uncertainty and variability in values and preferences*

We have noted that systematic study of patients' values and preferences are very limited. As a result, panels will often be uncertain about typical values and preferences. The greater is that uncertainty, the more likely they will make a weak recommendation.

Given the sparse systematic study of patients' values and preferences, one could argue that large uncertainty always exists about the patients' perspective. On the other hand, some systematic study of values and preferences and decision making has been completed, and clinicians' experience with patients may provide considerable additional insight.

Indeed, on occasion, panels will, on the basis of clinical experience, be confident regarding typical patient's values and preferences. Pregnant women's strong aversion to even a small risk of important fetal abnormalities may be one such situation [20].

A second concern that may make a weak recommendation more likely is large variability in values and preferences. To the extent large variability exists, it is less likely that a single recommendation would apply uniformly across all patients, and the right course of action is likely to differ between patients.

Empirical evidence may inform estimates of variability in recommendations. For instance, Devereaux et al. [23] asked patients at risk of atrial fibrillation how many serious gastrointestinal bleeds they would tolerate and still be willing to use an anticoagulant to prevent a stroke. Although most patients placed a high value on avoiding a stroke and were ready to accept a bleeding risk of 22% to reduce their chances of having a stroke by 8%, diversity in values and preferences was also apparent. A few patients were ready to accept only a small risk of bleeding to reduce their stroke risk by 8%. These data, consistent with other studies of values and preferences regarding anticoagulation in atrial fibrillation [18], suggest that only in patients at appreciable risk of stroke would a strong recommendation for warfarin be warranted.

Although systematic study will lead to the highest confidence, panelists may express confidence in their estimates of variability in values and preference on the basis of clinical experience. In the example cited earlier, clinicians may be confident not only that the typical expectant mother will have a strong aversion to even a small risk of important fetal abnormalities but also that these values and preferences are virtually uniform across the population.

On the other hand, clinical experience may leave a panel confident that values and preferences differ widely among patients. For example, clinical experience makes it clear that an expectant couples' desire to undergo a genetic test that increases the risk of spontaneous miscarriage will differ greatly depending on their willingness to act on knowledge about a fetal anomaly and their attitude toward the loss of a normal pregnancy. Situations such as these when recommendations are particularly dependent on differing values and preferences may dictate, in addition to making a weak recommendation, including descriptions of how varying values and preferences will determine the optimal decision [14].

A hopeful patient may place more emphasis on a small chance of benefit, whereas a pessimistic, risk-averse patient may place more emphasis on avoiding the risks associated with a potentially beneficial therapy. Some patients may

have a belief that even if the risk of an adverse event is low, they will be the person who will suffer such an adverse effect.

For example, in patients with idiopathic pulmonary fibrosis, evidence for the benefit of steroids warrants only low confidence, whereas we can be very confident of a wide range of adverse effects associated with steroids. The hopeful patient with pulmonary fibrosis may be enthusiastic about use of steroids, whereas the risk-averse patient is likely to decline.

2.3. Confidence in estimates of effect (quality of evidence)

Another determinant of the direction and strength of recommendations is our confidence in the estimates of effect. Typically, a strong recommendation is associated with high, or at least moderate, confidence in the effect estimates for critical outcomes. If one has high confidence for some critical outcomes (typically, benefits of an intervention), but low confidence for other outcomes considered critical (often long-term harms), then a weak recommendation is likely warranted. The more closely balanced the tradeoffs between desirable and undesirable outcomes, the more likely that low confidence for any critical outcome will result in a weak recommendation.

Even when an apparently large gradient exists in the balance of desirable vs. undesirable outcomes, panels will be appropriately reluctant to offer a strong recommendation if their confidence in effect estimates is low. This is in part because when confidence in the estimate of effect is lower, choice is more preference dependent.

For instance, the GRADE approach provides insight into how guideline panels should have handled the decision regarding hormone replacement therapy (HRT) in postmenopausal women in the 1990s when observational studies suggested a substantial reduction in cardiovascular risk [24] (which randomized trials subsequently proved false [25], at least in women appreciably past the menopause), and equally low quality evidence suggested an increase in the risk of breast cancer (which proved true [26]).

Guideline panels during the 1990s made recommendations that were presented, or at least interpreted, as strong recommendations. Many primary care physicians, responding to these recommendations, enthusiastically encouraged their postmenopausal patients to use HRT. Appropriately considering the lack of confidence in estimates, women with a low level of risk aversion might indeed have been inclined to use HRT. Those with a high level of risk aversion would, however, have declined HRT. Clearly, a weak recommendation for (or perhaps even against) HRT would have been warranted.

For some questions, investigators may not have directly measured critical outcomes (in particular quality of life). In such instances, even if surrogates are available, confidence in estimates is very likely to be low.

2.3.1. Low confidence in effect estimates may, rarely, be tied to strong recommendations

In general, we discourage guideline panels from making strong recommendations when their confidence in estimates of effect for critical outcomes is low or very low. We have identified five paradigmatic situations, however, in which strong recommendations may be warranted despite low or very low quality of evidence (Table 4). These situations can be conceptualized as ones in which a panel would have a low level of regret if subsequent evidence showed that their recommendation was misguided.

One paradigmatic situation occurs when panels have low confidence regarding the benefit of an intervention in a life or death situation. Consider patients suffering from life-threatening disseminated blastomycosis [27]. High quality evidence suggests that amphotericin is more toxic than itraconazole, and low quality evidence that it reduces mortality in this context. When considering the subpopulation of patients with life-threatening blastomycosis, panels may reason that all or virtually all patients would choose the more toxic therapy given the very high risk of death and the possibility that amphotericin may decrease that risk. If they did so, they would make a strong recommendation for amphotericin.

In a second paradigmatic situation, panels may make a strong recommendation against an intervention when there is uncertainty of benefits, but they are confident about adverse effects and resource use. For example, it remains very uncertain whether whole-body computed tomography scan or magnetic resonance imaging screening confers benefits in terms of reduction of cancer risk, but there is no doubt that such tests generate false positives that result in anxiety and possibly invasive tests with their own discomfort and complications [28]. Such tests also consume scarce resources. Despite the low confidence with regard to benefits, guideline panels might legitimately make strong recommendations against screening imaging.

A third situation occurs when we have low quality evidence regarding relative benefit, but high quality evidence of lower harm for one of the competing alternatives. For instance, in patients who have early-stage, low-grade, *Helicobacter pylori*−positive gastric mucosa-associated lymphoid tissue lymphoma, low quality evidence suggests that initial *H. pylori* eradication therapy results in similar rates of complete response (50−80%) in comparison with the alternatives of radiation therapy or gastrectomy [29]. The evidence warrants high confidence in the increased morbidity associated with either radiation or gastrectomy vs. pharmacologic therapy. Furthermore, in patients without complete response, there is the option of later use of the higher risk alternatives. Thus, despite low confidence in estimates of effects, a strong recommendation for *H. pylori* eradication therapy appears appropriate.

In a fourth situation, panels may make strong recommendations for one of the two competing alternatives if they are confident of similarity of benefits, but have only

Table 4. Paradigmatic situations in which a strong recommendation may be warranted despite low or very low confidence in effect estimates

Situation	Condition	Example
1	When low quality evidence suggests benefit in a life-threatening situation (evidence regarding harms can be low or high)	Fresh frozen plasma or vitamin K in a patient receiving warfarin with elevated INR and an intracranial bleed. Only low quality evidence supports the benefits of limiting the extent of the bleeding
2	When low quality evidence suggests benefit and high quality evidence suggests harm or a very high cost	Head-to-toe CT/MRI screening for cancer. Low quality evidence of benefit of early detection but high quality evidence of possible harm and/or high cost (strong recommendation against this strategy)
3	When low quality evidence suggests equivalence of two alternatives, but high quality evidence of less harm for one of the competing alternatives	*Helicobacter pylori* eradication in patients with early stage gastric MALT lymphoma with *H. pylori* positive. Low quality evidence suggests that initial *H. pylori* eradication results in similar rates of complete response in comparison with the alternatives of radiation therapy or gastrectomy; high quality evidence suggests less harm/morbidity
4	When high quality evidence suggests equivalence of two alternatives and low quality evidence suggests harm in one alternative	Hypertension in women planning conception and in pregnancy. Strong recommendations for labetalol and nifedipine and strong recommendations against angiotensin converting enzyme (ACE) inhibitors and angiotensin receptor blockers (ARB)—all agents have high quality evidence of equivalent beneficial outcomes, with low quality evidence for greater adverse effects with ACE inhibitors and ARBs
5	When high quality evidence suggests modest benefits and low/very low quality evidence suggests possibility of catastrophic harm	Testosterone in males with or at risk of prostate cancer. High quality evidence for moderate benefits of testosterone treatment in men with symptomatic androgen deficiency to improve bone mineral density and muscle strength. Low quality evidence for harm in patients with or at risk of prostate cancer

Abbreviations: INR, international normalized ratio; CT, computed tomography; MRI, magnetic resonance imaging; MALT, mucosa-associated lymphoid tissue.

low or very low confidence regarding increased harm for one alternative. Reasoning that there is nothing to lose, and possibly a lot to gain in terms of a lower incidence of adverse effects, guideline panels may reasonably make a strong recommendation for the agent apparently free from serious toxicity. For instance, consider the management of hypertension in women who are planning conception and who are pregnant. There is high quality evidence of equivalent effectiveness for labetalol, nifedipine, angiotensin-converting enzyme (ACE) inhibitors, and angiotensin receptor blockers (ARBs). There is low quality evidence of harms for ACE inhibitors and ARBs. Panels have appropriately made strong recommendations for labetalol and nifedipine and strong recommendations against ACE inhibitors and ARBs [30].

A fifth paradigmatic situation occurs when we have moderate-to-high confidence about an intervention's modest benefits, but remain uncertain about its likelihood of causing catastrophic harm. For example, high quality evidence supports the inference that testosterone is beneficial for men with symptomatic androgen deficiency, improving their quality of life and markers of bone and muscle strength. However, low quality evidence links testosterone use to an increased risk of prostate cancer. As a result, a panel of endocrinologists formulated a strong recommendation against testosterone use in men with prostate cancer and in men pending evaluation of palpable prostate nodule or induration or prostate-specific antigen (PSA) level of 4 ng/mL or PSA level of 3 ng/mL in men at high risk of prostate cancer [31].

2.4. Resource use

Panels may or may not consider resource use in their judgments about the direction and strength of recommendations. Reasons for not considering resource use include a lack of reliable data, the intervention is not useful and the effort of calculating resource use can be spared, the desirable effects so greatly outweigh any undesirable effects that resource considerations would not alter the final judgment, or they have elected (or been instructed) to leave resource considerations up to other decision makers.

Once again, panels should be explicit about the decision they made not to consider resource utilization and the reason for their decision. If they elect to include resource utilization when making a recommendation, but have not included resource use as a consequence when preparing an evidence profile, they should be explicit about what types of resource use they considered when making the recommendation and whatever logic or evidence was used in their judgments.

For example, a panel making a recommendation about oseltamivir for treatment of patients hospitalized with avian influenza (H5N1) in nonpandemic situations considered the cost of oseltamivir, but did not explicitly consider the quality of the evidence for resource use. Overall, the quality of

the underlying evidence for all recommendations was rated as very low because it was based on small case series of H5N1 patients, on extrapolation from preclinical studies, and high quality studies of seasonal influenza. A strong recommendation to treat H5N1 patients with oseltamivir was made in part because of the severity of the disease. With only very low quality evidence of the beneficial and adverse effects of oseltamivir for avian influenza, the panel decided not to consider quality of evidence for resource use. The panel summarized their thinking regarding resource use as a factor in making their recommendation by stating: ''The cost is not high for treatment of sporadic cases'' [32].

We discuss special challenges related to rating the confidence in estimates for resource use in another article in this series [9].

3. Special considerations of the determinants of direction and strength of recommendations

3.1. Baseline risk (control event rate) can influence the balance

Table 3 presents an example of how guideline panels can move from evidence to recommendations in an explicit and transparent way. The final column in Table 3 presents the issues (if one calls the four determinants domains, then one might call these issues subdomains) that guideline panels should consider under each domain. One of these subdomains, which may be critical in the decision, is baseline risk.

Because, we usually determine absolute risk differences through applying the relative risk reduction to a baseline risk [11], large baseline risk differences will result in large absolute risk differences. For example, recommendations for duration of anticoagulation in patients with deep venous thrombosis will differ depending on the likelihood of recurrent thrombosis. The likelihood of recurrent thrombosis differs in those with and without clear precipitating factors for the original thrombotic event—in particular, patients whose deep venous thrombosis is precipitated by a surgical procedure have a low risk of recurrence. Anticoagulation is associated with inconvenience and a risk of serious bleeding. Therefore, indefinite anticoagulation will seldom be appropriate in those at low risk of recurrence whose absolute benefit with anticoagulation is small, but may well be mandated in patients at much higher risk. Thus, the strength of recommendations—and likely the direction—will differ in high- and low-risk groups [33].

3.2. Recommendations may differ by setting and perspective

In our introductory discussion of globalizing evidence, localizing recommendations, we noted that we do not expect uniformity of recommendations across settings. Here, we expand the reasons for the anticipated diversity, and how differences in perspective can contribute.

The impact of an intervention may differ across geographic settings depending on the risk of adverse events in untreated population (e.g., risk of coronary events is much lower in low income countries), or the capacity to deliver the intervention (e.g., monitoring of anticoagulant therapy).

Values and preferences may differ among cultures, even if those cultures appear very similar. For example, after viewing the same evidence, American and New Zealand guideline developers came to different conclusions about the trade-offs associated with colon cancer screening [34−36].

Values may also differ in subcultures vs. mainstream culture within a population. For example, in formulating the CCIRH guidelines, the panel's awareness of immigrant populations' vulnerability to family disruption and possible deportation supported the recommendation against routine screening for intimate partner violence [37].

Finally, resource implications and opportunity cost may differ. For instance, a year's supply of an expensive drug may cost the equivalent of a single nurse's salary in the United States, 4 nurses' salaries in Poland, and 20 nurses' salaries in China.

In the face of the same evidence, recommendations may also differ according to perspective. Our discussion in this article has addressed, almost exclusively, guideline panels making recommendations from the perspective of patients and the health care providers looking after those patients. Sometimes, however, a panel may make recommendations from a public health or societal perspective.

For example, panels making recommendations about H1N1, avian, or seasonal influenza may place a large value on outcomes that may not be directly critical or important to individual patients, such as reducing the spread of disease [32,38]. Other times, a panel may make recommendations from the perspective of the government or a private insurance company, placing a large value on costs (or alternative uses of resources) within a fixed budget. Equity, feasibility, and burden of illness may be other considerations important to public policy decision making, but of much less relevance to individual decision making. Panels should explicitly state the perspective they are taking, particularly when they are not taking a patient-centered perspective.

3.3. Evidence to recommendations synthesis

As in Table 3, GRADE suggests that guideline panels present a synthesis of their judgments about the domains determining direction and strength of recommendations, and how this synthesis informs the recommendation. Disagreement between panels is common [39−41], and disagreement may be a result of variability in judgments about the domains or of how panels synthesize those judgments. Presentation and publication of frameworks

summarizing the rationale for recommendations can support transparency in the decision process and be used for stakeholder engagement (Table 3).

Consider, for example, views expressed in the literature concerning the merits of perioperative use of beta-blockers in patients undergoing noncardiac surgery. Some assert that lower doses of beta-blockers administered well before surgery could prevent the documented increase in stroke risk with beta-blockers [42,43]. Others do not agree [44]. An evidence to action synthesis from the former group would emphasize the heterogeneity of results from trials that used different doses and different periods of administration of beta-blockers before surgery, and the latter would not.

Alternatively, disagreement in recommendations might be because they have different views of the relative value of reducing the risk of MI with beta-blocker use (approximately 1.5% in those at 5% baseline risk) vs. the increase in stroke risk (approximately 0.5% in those at 0.5% baseline risk of stroke). Both may agree that patients value preventing stroke more than preventing MI, but the synthesis from a panel recommending against beta-blockers would emphasize that the patients generally place very high value in avoiding disabling stroke and the asymptomatic nature of many perioperative MIs.

4. Conclusion

Patients, clinicians, and policy makers will all be better served by a more systematic and transparent system for judging the direction and strength of recommendations. Explicit presentation of how panels view the four domains to consider in the direction and strength of recommendations could play an important role in improving the transparency of panel decisions (Table 3).

References

[1] Guyatt G, Oxman A, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64:395−400.

[2] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401−6.

[3] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol 2011;64:407−15.

[4] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence-publication bias. J Clin Epidemiol 2011;64:1277−82.

[5] Guyatt G, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. J Clin Epidemiol 2011;64:1283−93.

[6] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence-inconsistency. J Clin Epidemiol 2011;64:1294−302.

[7] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence-indirectness. J Clin Epidemiol 2011;64:1303−10.

[8] Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011;64:1311−6.

[9] Brunetti M, Shemilt I, Pregno S, Vale L, Oxman A, Lord J, et al. GRADE guidelines: 10. Considering resource use and rating the quality of economic evidence. J Clin Epidemiol 2013;66:140−50.

[10] Guyatt GH, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol 2013;64:151−7.

[11] Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables: binary outcomes. J Clin Epidemiol 2013;66:158−72.

[12] Guyatt GH, Thorlund K, Oxman AD, Walter S, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables: continuous outcomes. J Clin Epidemiol 2013;66:173−83.

[13] Guyatt G, Oxman A, Akl E, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383−94.

[14] Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol 2013;66:719−25.

[15] Eisenberg JM. Globalize the evidence, localize the decision: evidence-based medicine and international diversity. Health Aff (Millwood) 2002;21(3):166−8.

[16] Schunemann HJ, Woodhead M, Anzueto A, Buist S, Macnee W, Rabe KF, et al. A vision statement on guideline development for respiratory disease: the example of COPD. Lancet 2009;373:774−9.

[17] Montori V, Devereaux P, Straus S, Haynes B, Guyatt G. Decision making and the patient. In: Guyatt G, editor. The users' guides to the medical literature: a manual for evidence-based clinical practice. 2nd ed. New York, NY: McGraw-Hill; 2008.

[18] McLean S, Mulla S, Akl EA, Jankowski M, Vandvik P, Ibrahim S, et al. Patient values and preferences in decision making for antithrombotic therapy: a systematic review. Antithrombotic therapy and prevention of thrombosis, 9th ed.: American College of Chest physicians Evidence-Based Clinical Practice Guidelines. Chest 2012;141(2 Suppl):e1S−23S.

[19] Pottie K, Greenaway C, Feightner J, Welch V, Swinkels H, Rashid M, et al. Evidence-based clinical guidelines for immigrants and refugees. CMAJ 2011;183(12):E824−925.

[20] Bates S, Greer I, Middeldorp S, Veenstra D, Prabulos A, Vandvik P, et al. VTE, thrombophilia, antithrombotic therapy, and pregnancy: Antithrombotic therapy and prevention of thrombosis, 9th ed.: American College of Chest physicians Evidence-Based Clinical Practice Guidelines. Chest 2012;141(2 Suppl):e691S−736S.

[21] Goodman SG, Menon V, Cannon CP, Steg G, Ohman EM, Harrington RA. Acute ST-segment elevation myocardial infarction: American College of Chest physicians evidence-based clinical practice guidelines (8th Edition). Chest 2008;133(6 Suppl):708S−75S.

[22] Webster A, Woodroffe R, Taylor R, Chapman J, Craig J. Tacrolimus versus cyclosporin as primary immunosuppression for kidney transplant recipients. Cochrane Database Syst Rev 2006;4:CD003961.

[23] Devereaux PJ, Anderson DR, Gardner MJ, Putnam W, Flowerdew GJ, Brownell BF, et al. Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study. BMJ 2001;323:1218−22.

[24] Guidelines for counseling postmenopausal women about preventive hormone therapy. American College of Physicians. Ann Intern Med 1992;117:1038−41.

[25] 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. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. JAMA 1998;280:605−13.

[26] Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. 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.

[27] Chapman SW, Dismukes WE, Proia LA, Bradsher RW, Pappas PG, Threlkeld MG, et al. Clinical practice guidelines for the management of blastomycosis: 2008 update by the Infectious Diseases Society of America. Clin Infect Dis 2008;46:1801−12.

[28] Lauenstein TC, Semelka RC. Emerging techniques: whole-body screening and staging with MRI. J Magn Reson Imaging 2006; 24(3):489−98.

[29] Malfertheiner P, Megraud F, O'Morain C, Bazzoli F, El-Omar E, Graham D, et al. Current concepts in the management of Helicobacter pylori infection: the Maastricht III Consensus Report. Gut 2007;56(6):772−81.

[30] Magee LA, Helewa M, Moutquin JM, van Dadelszen P, for the Hypertension Guideline Committee. Diagnosis, evaluation, and management of the hypertensive disorders of pregnancy. SOGC Clinical Practice Guideline, No. 206, March 2008. J Obstet Gynaecol Can 2008;30: S1−48.

[31] Bhasin S, Cunningham G, Hayes F, Matsumoto A, Snyder P, Swerdloff R, et al. Testosterone therapy in men with androgen deficiency syndromes: an Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 2010;95:2536−59.

[32] Schunemann HJ, Hill SR, Kakad M, Bellamy R, Uyeki TM, Hayden FG, et al. WHO Rapid Advice Guidelines for pharmacological management of sporadic human infection with avian influenza A (H5N1) virus. Lancet Infect Dis 2007;7:21−31.

[33] Kearon C, Akl EA, Comerota AJ, Prandoni P, Bounameaux H, Goldhaber SZ, et al. Antithrombotic therapy for VTE disease: antithrombotic therapy and prevention of thrombosis, 9th ed.: American College of Chest physicians Evidence-Based Clinical Practice Guidelines. Chest 2012;141(2 Suppl):e419S−94S.

[34] Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2008;149:627−37.

[35] Levin B, Lieberman DA, McFarland B, Andrews KS, Brooks D, Bond J, et al. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. Gastroenterology 2008;134(5):1570−95.

[36] Guidance on surveillance for people at increased risk of colorectal cancer. Wellington, New Zealand: New Zealand Guidelines Group; 2012.

[37] Tugwell P, Pottie K, Welch V, Ueffing E, Chambers A, Feightner J. Evaluation of evidence-based literature and formulation of recommendations for the clinical preventive guidelines for immigrants and refugees in Canada. CMAJ 2011;183(12):E933−8.

[38] Schunemann HJ, Hill SR, Kakad M, Vist GE, Bellamy R, Stockman L, et al. Transparent development of the WHO rapid advice guidelines. PLoS Med 2007;4(5):e119.

[39] Oxman AD, Glasziou P, Williams JW Jr. What should clinicians do when faced with conflicting recommendations? BMJ 2008;337:a2530.

[40] Georg G, Colombet I, Durieux P, Menard J, Meneton P. A comparative analysis of four clinical guidelines for hypertension management. J Hum Hypertens 2008;22(12):829−37.

[41] Matthys J, De Meyere M, van Driel ML, De Sutter A. Differences among international pharyngitis guidelines: not just academic. Ann Fam Med 2007;5(5):436−43.

[42] Kaafarani HM, Atluri PV, Thornby J, Itani KM. beta-Blockade in noncardiac surgery: outcome at all levels of cardiac risk. Arch Surg 2008;143:940−4. discussion 944.

[43] van Lier F, Schouten O, van Domburg RT, van der Geest PJ, Boersma E, Fleisher LA, et al. Effect of chronic beta-blocker use on stroke after noncardiac surgery. Am J Cardiol 2009;104:429−33.

[44] Bangalore S, Wetterslev J, Pranesh S, Sawhney S, Gluud C, Messerli FH. Perioperative beta blockers in patients having noncardiac surgery: a meta-analysis. Lancet 2008;372:1962−76.