

GRADE SERIES

GRADE guidelines: 14. Going from evidence to recommendations:  
the significance and presentation of recommendations

Jeff Andrews<sup>a,\*</sup>, Gordon Guyatt<sup>b,c,d</sup>, Andrew D. Oxman<sup>e</sup>, Phil Alderson<sup>f</sup>, Philipp Dahm<sup>g</sup>,  
Yngve Falck-Ytter<sup>h</sup>, Mona Nasser<sup>i</sup>, Joerg Meerpohl<sup>j,k</sup>, Piet N. Post<sup>l</sup>, Regina Kunz<sup>m</sup>,  
Jan Brozek<sup>b,c</sup>, Gunn Vist<sup>e</sup>, David Rind<sup>n,o</sup>, Elie A. Akl<sup>p</sup>, Holger J. Schünemann<sup>b,c,d</sup>

<sup>a</sup>Vanderbilt Evidence-based Practice Center, Vanderbilt University, #27166-719 Thompson Lane, Nashville, TN 37204-3195, USA

<sup>b</sup>Department of Clinical Epidemiology, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

<sup>c</sup>Department of Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

<sup>d</sup>Department of Medicine, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

<sup>e</sup>Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs plass, Oslo 0130, Norway

<sup>f</sup>National Institute for Health and Clinical Excellence, Level 1A, City Tower, Piccadilly Plaza, Manchester M1 4BD, UK

<sup>g</sup>Department of Urology, College of Medicine, University of Florida and VA Medical Center, Box 100247, Room N2-15, Health Science Center, Gainesville, FL 32610, USA

<sup>h</sup>Division of Gastroenterology, Case and VA Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA

<sup>i</sup>Peninsula College of Medicine and Dentistry, Universities of Exeter and Plymouth, The John Bull Building, Tamar Science Park, Plymouth PL68BU, UK

<sup>j</sup>German Cochrane Center, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Freiburg 79110, Germany

<sup>k</sup>Pediatric Hematology and Oncology, Center for Pediatrics and Adolescent Medicine, University Medical Center Freiburg, Freiburg 79106, Germany

<sup>l</sup>Post Voor Zorg, Van Barenstraat 31, 2628 LC Delft, The Netherlands

<sup>m</sup>asim, Academy of Swiss Insurance Medicine, University Hospital Basel, Petersgraben 4, Basel 4031, Switzerland

<sup>n</sup>Harvard Medical School, Boston, MA, USA

<sup>o</sup>UpToDate, Boston, MA, USA

<sup>p</sup>Department of Medicine, State University of New York at Buffalo, NY, USA

Accepted 12 March 2012; Published online 9 January 2013

Abstract

This article describes the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach to classifying the direction and strength of recommendations. The strength of a recommendation, separated into strong and weak, is defined as the extent to which one can be confident that the desirable effects of an intervention outweigh its undesirable effects. Alternative terms for a weak recommendation include conditional, discretionary, or qualified. The strength of a recommendation has specific implications for patients, the public, clinicians, and policy makers. Occasionally, guideline developers may choose to make “only-in-research” recommendations. Although panels may choose not to make recommendations, this choice leaves those looking for answers from guidelines without the guidance they are seeking. GRADE therefore encourages panels to, wherever possible, offer recommendations. © 2013 Published by Elsevier Inc.

Keywords: GRADE; Quality of evidence; Strength of evidence; Guideline development; Grading; Recommendations

1. Introduction

In prior papers in this series devoted to the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach to systematic reviews and practice

guidelines, we have dealt with the process before developing recommendations: framing the question [1], choosing critical and important outcomes [2], rating the confidence in effect estimates for each outcome [3–9], rating the confidence in effect estimates across outcomes [10], dealing with resource use [11], creating an evidence profile and a Summary of Findings (SoF) table [12,13], and GRADE’s approach to diagnostic test recommendations. This article addresses GRADE’s approach to categorizing, labeling, and wording health care recommendations. As we did in the initial article in this series, we will define strong or

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.

\* Corresponding author.

E-mail address: rafterl@mcmaster.ca (J. Andrews).

## What is new?

### Key points

- The strength of a recommendation is defined as the extent to which one can be confident that the desirable consequences of an intervention outweigh its undesirable consequences.
- Grading of Recommendations Assessment, Development, and Evaluation GRADE has chosen a simple four-category classification of recommendations, a binary classification of recommendations as strong or weak (also known as conditional, discretionary, or qualified) recommendations for or against a management approach.
- The strength of a recommendation has specific implications for patients, the public, clinicians, and policy-makers.

weak recommendations for or against a particular management approach, and discuss the interpretation and presentation of these recommendations. In the next article in the series, we will focus on the process of going from the evidence to the recommendations. Throughout this article, we will refer to guideline developers as “the panel.”

## 2. Presenting direction and strength of recommendations

### 2.1. Direction of recommendations

Panels make recommendations either for (when the desirable consequences outweigh the undesirable consequences) or against (when the opposite is true) a particular strategy, in relation to a comparator. With the GRADE approach, the desirable and undesirable consequences are the outcomes classified as “critical” and “important but not critical.” These outcomes are selected at the outset, confirmed when the results are reviewed, and presented in the evidence profile and SoF table.

In almost all situations, there are trade-offs between management strategies that have some desirable and some undesirable outcomes. Table 1 presents typical categories of desirable and undesirable consequences of a management strategy. Inevitably, evaluating the balance between desirable and undesirable consequences involves judging the relative importance of those consequences, an issue we will address in the next article.

### 2.2. Strength of recommendations

Like confidence in effect estimates (quality of evidence), the strength of a recommendation can be conceptualized as

**Table 1.** Categories of typical desirable and undesirable outcomes of an experimental vs. a control intervention

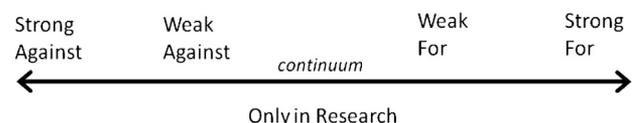
| Desirable outcomes  | Undesirable outcomes   |
|---|--|
| <ul style="list-style-type: none"> <li>• Increase longevity</li> <li>• Reduction in morbid events intervention designed to prevent</li> <li>• Resolution of symptoms</li> <li>• Improved quality of life</li> <li>• Decreased resource use</li> </ul> | <ul style="list-style-type: none"> <li>• Decreased longevity</li> <li>• Immediate serious complications (typically for surgical therapies)</li> <li>• Short-term relatively minor side effects</li> <li>• Long-term rare serious adverse events</li> <li>• Impaired quality of life</li> <li>• Inconvenience/hassle</li> <li>• Increased resource use</li> </ul> |

an underlying continuum (Fig. 1). Nevertheless, GRADE has chosen a simple four-category classification of recommendations. If the panel is highly confident of the balance between desirable and undesirable consequences, they make a strong recommendation for (desirable outweighs undesirable) or against (undesirable outweighs desirable) an intervention. If the panel is less confident of the balance between desirable and undesirable consequences, they offer a weak recommendation. Some panels have been concerned about the use of “weak” to characterize recommendations because a weak recommendation can be confused with weak evidence, because guideline users may feel they can ignore weak recommendations, or because users may interpret weak as denoting that the panel was uncertain regarding the right recommendation. GRADE therefore offers alternative labels: conditional, discretionary, and qualified (Box 1) [14]. As we will demonstrate, the four-category approach to grading recommendations has the merit not only of simplicity, but also of direct links to action on the part of health care providers, health care recipients, and policy makers.

### 2.3. Presentation of recommendations

Recommendations in the passive voice may lack clarity. We therefore suggest that guideline developers present recommendations in the active voice. For example, a number of organizations use “we recommend...” and “we suggest...” for strong and weak recommendations, respectively. Alternatives for a strong recommendation are

#### Strength of recommendation on a continuum: categorical terminology



**Fig. 1.** Strength of recommendation: a continuum divided into categories.

**Box 1 Terminology: weak recommendations**

We have referred to recommendations as strong and weak. However, some guideline panels experience an unintended negative connotation with the word “weak,” and possible unintended conflation with “weak evidence.” We suggest three alternative terms that panels may choose to use: conditional, discretionary, or qualified. Recommendations may be conditional upon patient values and preferences, the resources available or the setting in which the intervention will be implemented. Recommendations may be at the discretion of the patient and clinician, or qualified with an explanation about the issues that would lead decisions to vary.

“Clinicians should...” or “Clinicians should not...” or “Do...” or “Don’t...” Alternatives for a weak recommendation include “Clinicians might...” or “We conditionally recommend...” or “We make a qualified recommendation that...” (Box 1).

There is, however, limited systematically collected evidence addressing the wording of the strength of recommendations. In a randomized trial, we compared three wording approaches that expressed two grades of recommendation (“we recommend”/“we suggest”; “clinicians should”/“clinicians might”; “we recommend”/“we conditionally recommend”) [15]. None of the approaches was clearly superior to the others in conveying the strength of recommendations. Lomotan et al. [16] compared the “level of obligation” assigned to various terms commonly used in health care guidelines. They found that participants assigned different levels of obligation to “must,” “should,” and “may.”

Recommendations should always specify the population, and unless it is obvious, the comparator. Consider for instance, the following: In patients with acute renal failure, we recommend hourly urine volume measurement for at least 24 hours. The strength of this recommendation may differ depending on whether the alternative is every 2 hours or once a day. Thus, the additional specification “when compared with daily urine volume measurement” is required.

Sometimes, the recommendation statement will include reference to the setting, particularly when our confidence in estimates of effect would vary according to the setting. For instance, a recommendation regarding carotid endarterectomy might vary depending on the extent of delay between a patient’s presentation with symptoms suggesting carotid stenosis and the performance of surgery [17]. Another instance when setting may be important is an expensive intervention in high- vs. low-income countries.

In general, it is preferable to present recommendations in favor of a particular management approach rather than

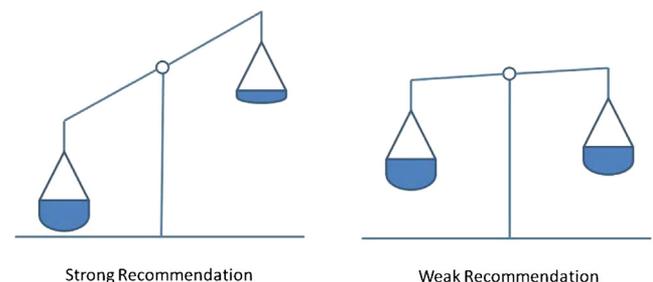
against an approach. For instance, in considering the addition of aspirin to clopidogrel in patients who have had a stroke, it would be preferable to state: “In patients who have had a stroke, we suggest clopidogrel alone vs. adding aspirin to clopidogrel” rather than “In patients who have had a stroke and are using clopidogrel, we suggest not adding aspirin.”

Nevertheless, when a useless or harmful therapy is in wide use, recommendations against a management approach are appropriate. For instance, “In patients undergoing cardiac surgery who were not previously receiving beta blockers, we suggest not initiating perioperative beta blocker therapy.”

Unfortunately, misinterpretation is possible however strength of recommendations is expressed. We suggest guideline developers consider using both symbols (which may be less confusing than numbers or letters [18]) and words to express strength of recommendations. We suggest  $\uparrow\uparrow$  as a symbol for strong recommendations and  $\uparrow?$  for weak recommendations. For guideline developers preferring numbers or letters, we suggest “1” for strong recommendations and “2” for weak. For those who prefer a pictorial representation, balancing scales are depicted in (Fig. 2). Whatever terms guideline developers elect to use (e.g., weak, conditional, discretionary, or qualified), we suggest that they use these consistently across different guidelines. Explanations of the meaning and implications of strong and weak recommendations should be readily accessible, for example, using hyperlinks in electronic publications, to facilitate correct interpretation.

**3. Meaning of recommendations in GRADE***3.1. What GRADE means by strong and weak recommendations—for clinicians and patients*

Using the GRADE approach, guideline authors make a strong recommendation when they believe that all or almost all informed people would make the recommended choice for or against an intervention. Consider, for example, the recommendation to take supplemental folate before and during the pregnancy. High-quality evidence suggests folate prevents neural tube defects, a catastrophic outcome of pregnancy [19,20]. Folate is inexpensive and



**Fig. 2.** Balance scales to depict strong vs. weak recommendations.

has no proven adverse effects. Because the desirable consequences so greatly outweigh the negative, the deduction that all informed women would choose to take supplemental folate is secure, thus warranting a strong recommendation.

In contrast, guideline panels using GRADE make a weak recommendation when they believe that most informed people would choose the recommended course of action, but a substantial number would not. Consider the recommendation in favor of adjuvant chemotherapy for women with early stage breast cancer. Most women would choose the recommended course of action, but an appreciable number would choose not to take chemotherapy, because they feel that the small possible benefits in survival do not justify the suffering resulting from the serious side effects of chemotherapy [21].

Given that a strong recommendation implies uniformity of choice and a weak recommendation implies variability, strong and weak recommendations have direct implications for the patient–provider dyad at the point of decision making. Although recognizing that it is always valuable for providers to discuss decisions with patients, allocation of time will differ given the strength of a recommendation. When a recommendation is weak, clinicians and other health care providers need to devote more time to the process of shared decision making by which they ensure that the informed choice reflects individual values and preferences (Box 1). This is likely to involve ensuring patients understand the implications of the choices they are making, possibly using a formal decision aid. When recommendations are strong, clinicians may spend less time on the process of making a decision, and focus efforts on overcoming barriers to implementation or adherence.

### 3.2. What GRADE means by strong and weak recommendations—for policy makers

The implication of a strong recommendation for policy makers is that the recommendation can be adopted as a policy in most situations. A strong recommendation implies that variability in clinical practice between individuals or regions would likely be inappropriate. Thus, for governments, institutions, provider groups, or third-party payers responsible for ensuring high-quality care, strong recommendations also constitute candidates for performance measures (quality of care criteria). For policy makers, the implication of a weak recommendation is that policy making will require substantial debate and involvement of many stakeholders. A weak recommendation implies that variability between individuals or regions may be appropriate, and use as a quality of care criterion is inappropriate unless the criterion is whether patients were properly informed and helped to make a decision consistent with their own values (such as by the use of a decision aid).

### 3.3. Strong does not necessarily mean a priority recommendation

The strength of a recommendation may not be directly correlated with its priority for implementation. The importance or prioritization of a recommendation may differ, depending on the target audience for the recommendation: patients, the public, clinicians, or policy makers. **Governments and public health officials considering a public health intervention must consider several factors beyond the strength of a recommendation.** These factors—of lesser relevance to recommendations directed at clinicians—include **the prevalence of the health problem** (higher priority for more common conditions), **ease of implementation** (higher priority for interventions that can be implemented now), **considerations of equity** (higher priority for interventions that contribute to reducing address health inequities), **total costs to society** (lower priority for interventions with high total costs), and **the potential for improvement in quality of care** (lower priority for recommendations with current high adherence). Therefore, government and public health officials may place a lower priority on implementing strong recommendations although they are important for individual patients. For instance, a National Institute for Clinical Excellence (NICE) guideline concerning hip fractures did not consider implementation of a recommendation to use an intramedullary nail in patients with subtrochanteric fracture a high priority because the practice is already widespread [22].

If guideline panels are addressing funders or health system managers, they should make transparent the manner in which factors related to prevalence, equity, cost, and improving quality of care influence their priorities. Sometimes these same factors can influence recommendations, particularly when guideline panels are making recommendations for clinicians and patients on behalf of funders. When this is the case, they should be explicit about the additional factors that are considered, this should be done consistently, and it should be transparent when these other factors influenced a recommendation.

## 4. Transparent values and preferences

In this section, we deal with the explicit and transparent presentation of the values and preferences underlying recommendations (Box 2). In the next article in the series, we deal with the sources of the values and preferences and how to use them in the process of making recommendations.

Ideally, guidelines will state foundational assumptions about the values and preferences that underlie their recommendations for the target population. For instance, a guideline addressing issues of thrombosis prevention and treatment in pregnancy noted: “Our recommendations reflect a belief that most women will place a low value on avoiding the pain, cost, and inconvenience of heparin

### Box 2 Terminology: values and preferences

Values and preferences is an overarching term that includes patients' perspectives, beliefs, expectations, and goals for health and life [37]. 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.

therapy to avoid the small risk of even a minor abnormality in their child" associated with warfarin prophylaxis [23].

In addition to, or in place of, making such general statements, panels may find it appropriate to make statements associated with specific recommendations that are particularly sensitive to values and preferences. For instance, two panels that were part of a broader guideline effort made apparently contradictory recommendations regarding aspirin vs. clopidogrel in patients with atherosclerotic vascular disease, despite using the same underlying evidence from a trial that enrolled both patients with threatened stroke and those with peripheral vascular disease [24]. The stroke panel that recommended clopidogrel over aspirin stated: "This recommendation... places a relatively high value on a small absolute risk reduction in stroke rates, and a relatively low value on minimizing drug expenditures [25]." The peripheral vascular disease panel that recommended aspirin over clopidogrel, stated: "This recommendation places a relatively high value on avoiding large expenditures to achieve small reductions in vascular events" [26]. The recommendations suggest opposite courses of action. Both are appropriate given the stated values and preferences, which were made explicit in qualifying statements accompanying each recommendation. These conflicting recommendations illustrate the importance of the values and preferences underlying the recommendations, the source of which we will discuss in the next article.

Another way to frame values and preferences statements that panels may want to consider is in terms of patients who do not share the values and preferences underlying the recommendation. UpToDate uses this approach. For instance, in their topic dealing with the treatment of achalasia they say: "For most healthy patients undergoing an invasive procedure, we suggest minimally invasive surgical myotomy rather than pneumatic dilatation. Patients who prefer to avoid surgery and the high rates of gastroesophageal reflux disease seen after surgery, and who are willing to accept a higher initial failure rate and long-term recurrence rate, can reasonably choose pneumatic dilatation" [27].

The text describing the rationale for the recommendations should state which outcomes the panel judged critical, which important, and which were not included. For recommendations particularly dependent on values and preferences, and those for which values and preferences are less certain, authors should place statements about underlying values and preferences with the recommendation statement rather than in the accompanying text.

For instance, a guideline panel made a recommendation for thrombolytic therapy in the context of acute stroke [28]. Thrombolytic therapy improves long-term functional outcome at the cost of an increase in immediate bleeding that is sometimes fatal. Thus, the panel felt compelled to add the following statement immediately following the recommendation: "This recommendation places relatively more weight on overall prospects for long-term functional improvement despite the increased risk of symptomatic intracerebral hemorrhage in the immediate peristroke period." This prominent positioning of the statements will make it less likely that consumers of the guidelines miss the importance of the values and preference judgments.

## 5. Special recommendation in GRADE

### 5.1. Recommendations to use interventions only in research may be appropriate

Panels may face decisions about promising interventions associated with appreciable harms or costs and with insufficient evidence of benefit to support their use. They may be reluctant, on one hand, to recommend against such interventions out of fear that they will stifle further investigation. At the same time, they may worry about encouraging the rapid diffusion of potentially ineffective or harmful interventions, and preventing recruitment to research already under way, by providing premature favorable recommendations for their use.

The adverse consequences of recommendations to use diethylstilbestrol for the prevention of miscarriage [29,30] highlight the risk of premature favorable recommendations (risks in the children of clear cell adenocarcinoma of the vagina and cervix, breast cancer, reproductive tract anomalies, infertility, and undescended testicles). When interventions have a large component of fixed costs such as equipment or facilities, an additional problem with premature recommendations in favor of an intervention is the risk of irretrievable allocation of resources that would be better spent elsewhere. Consider, for instance, the impact of prior recommendations to use continuous electronic fetal heart rate monitoring during labor in low-risk pregnancy [31,32].

Recommendations for use of an intervention only in the context of research may ameliorate these problems. Such a recommendation may provide an important stimulus to efforts to answer important research questions, thus

resolving uncertainty about optimal patient management [33]. For instance, a NICE guideline addressing management of patients with hip fracture noted the lack of a clear management pathway for patients admitted from care homes, the lack of randomized trials, and identified this as a research priority [22].

Only-in-research recommendations will be appropriate when three conditions are met: there is insufficient evidence supporting an intervention for a panel to recommend its use; further research has a large potential for reducing uncertainty about the effects of the intervention; and further research is deemed good value for the anticipated costs.

The research recommendations should be detailed regarding the specific research questions that investigators should address, particularly which patient-important outcomes they should measure [34]. The recommendation for research may be accompanied by an explicit strong recommendation not to use the experimental intervention outside of the research context.

### 5.2. Guideline panels may choose to not make recommendations

Not infrequently, panels may find themselves reluctant to make a recommendation for or against a particular management strategy, and also conclude that an “only-in-research” recommendation is inappropriate. There are two very different reasons for reluctance to make recommendations. One is that the confidence in effect estimates is so low that the panels feel a recommendation is too speculative. The US Preventative Services Task Force (USPSTF) has provided a thoughtful discussion of this situation, and some compelling examples (e.g., visual inspection to screen for skin cancer) [35].

The second reason is that although our confidence in effect estimates is moderate or even high, the trade-offs are so closely balanced, and the values and preferences and resource implications not known or too variable, that the panel has great difficulty deciding on the direction of a recommendation.

The USPSTF has remarked that clinicians “indicate frustration with the lack of guidance” when the task force fails to make recommendations [35]. As the USPSTF states: “Decision makers do not have the luxury of waiting for certain evidence. Even though evidence is insufficient, the clinician must still provide advice, patients must make choices, and policy makers must establish policies [35].”

Clinicians will rarely explore the evidence as thoroughly as a guideline panel, nor devote as much thought to the trade-offs, or the possible underlying values and preferences in the population. We therefore encourage panels to deal with their discomfort and to make recommendations even when confidence in effect estimate is low and/or desirable and undesirable consequences are closely balanced. Such recommendations will inevitably be weak, and may be accompanied by qualifications.

In the unusual circumstances in which panels choose not to make recommendations, they should specify whether this is on the basis of very low confidence in effect estimates, or because they feel the balance between desirable and undesirable consequences is so close they cannot make a recommendation.

A third reason a panel may be reluctant to make a recommendation is that two management options have very different undesirable consequences, and individual patients’ reactions to these consequences are likely to be so different that it makes little sense to think about typical values and preferences. Consider, for instance, adult patients with thalassemia major considering hematopoietic cell transplantation vs. continued medical treatment with transfusion and iron chelation. Such patients may face, on one hand, a possibility of cure of their thalassemia with transplant but an early mortality risk of approximately 33%, and on the other the prospect of continued morbidity and an uncertain prognosis. A guideline panel may consider that in such situations, the only sensible recommendation is a discussion between patient and physician to ascertain the patient’s preferences. Guideline panels should not, however, fail to make a recommendation simply because individual patients will make differing choices: that patients will make differing choices is a defining feature of a weak recommendation.

## 6. Conclusion

Guideline developers have used widely varying presentations of recommendations, and generally fail to specify the implications of recommendations for patients, clinicians, and policy makers. For instance, Hussain et al. [36] observed important variation in formulations of recommendations within and across guidelines. GRADE’s approach to standardized terminology and presentation, and clear specification of the implications of strong and weak recommendations, addresses these shortcomings.

## References

- [1] Guyatt GH, Oxman AD, 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] Guyatt G, Oxman AD, Akl EA, 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.
- [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] Balshem H, Helfand M, Schunemann H, Oxman AD, Kunz R, Brozek J, et al. Grade guidelines: 3 Rating the quality of evidence—introduction. *J Clin Epidemiol* 2011;64:401–6.
- [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, 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.

- [7] 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.
- [8] 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.
- [9] 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.
- [10] Guyatt GH, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 11. Making an overall rating of the quality of evidence for a single outcome and for all outcomes. *J Clin Epidemiol* 2013;66:151–7.
- [11] Brunetti M, Shemilt I, Pregno S, Vale L, Oxman AD, Lord J, et al. GRADE guidelines 11. Special challenges: confidence in estimates for resource use. *J Clin Epidemiol* 2013;66:140–50.
- [12] 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.
- [13] 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.
- [14] Chong L, Nasser M, Glasziou P. What should we call weak recommendations? *News Int Soc Evid Based Health Care* 2011;2:6.
- [15] Akl E, Guyatt GH, Levine M, Feldstein D, Irani J, Shaneyfelt T, et al. “Might” or “suggest”? No wording approach was clearly superior in conveying the strength of recommendation. *J Clin Epidemiol* 2012;65:268–75.
- [16] Lomotan EA, Michel G, Lin Z, Shiffman RN. How “should” we write guideline recommendations? Interpretation of deontic terminology in clinical practice guidelines: survey of the health services community. *Qual Saf Health Care* 2010;19(6):509–13.
- [17] Rothwell PM. External validity of randomised controlled trials: “to whom do the results of this trial apply?”. *Lancet* 2005;365:82–93.
- [18] Akl EA, Maroun N, Guyatt G, Oxman AD, Alonso-Coello P, Vist GE, et al. Symbols were superior to numbers for presenting strength of recommendations to health care consumers: a randomized trial. *J Clin Epidemiol* 2007;60:1298–305.
- [19] Folic acid for the prevention of neural tube defects: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med* 2009;150:626–31.
- [20] Wolff T, Witkop CT, Miller T, Syed SB. Folic acid supplementation for the prevention of neural tube defects: an update of the evidence for the U.S. Preventive Services Task Force. *Ann Intern Med* 2009;150:632–9.
- [21] Whelan T, Sawka C, Levine M, Gafni A, Reyno L, Willan A, et al. Helping patients make informed choices: a randomized trial of a decision aid for adjuvant chemotherapy in lymph node-negative breast cancer. *J Natl Cancer Inst* 2003;95:581–7.
- [22] National Institute for Health and Clinical Excellence. Hip fracture: the management of hip fracture in adults. Clinical guideline 124. London, UK; National Institute for Health and Clinical Excellence; June 2011.
- [23] Bates SM, Greer IA, Pabinger I, Sofaer S, Hirsh J. Venous thromboembolism, thrombophilia, antithrombotic therapy, and pregnancy: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th edition). *Chest* 2008;133(6 Suppl):844S–86S.
- [24] A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). CAPRIE Steering Committee. *Lancet* 1996;348:1329–39.
- [25] Albers GW, Amarencu P, Easton JD, Sacco RL, Teal P. Antithrombotic and thrombolytic therapy for ischemic stroke: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. *Chest* 2004;126(3 Suppl):483S–512S.
- [26] Clagett GP, Sobel M, Jackson MR, Lip GY, Tangelder M, Verhaeghe R. Antithrombotic therapy in peripheral arterial occlusive disease: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. *Chest* 2004;126(3 Suppl):609S–26S.
- [27] Spechler SJ, Achalasia. In: UpToDate, Grover S, deputy editor, Basow DS, Editor. Waltham, MA; UpToDate; April 25, 2012.
- [28] Albers GW, Amarencu P, Easton JD, Sacco RL, Teal P. Antithrombotic and thrombolytic therapy for ischemic stroke: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th edition). *Chest* 2008;133(6 Suppl):630S–69S.
- [29] Apfel RJ, Fisher SM. To do no harm: DES and the dilemmas of modern medicine. New Haven, CT: Yale University Press; 1984.
- [30] Dutton DB. Worse than the disease: pitfalls of medical progress. Cambridge, UK: Cambridge University Press; 1988.
- [31] Alfirevic Z, Devane D, Gyte G. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. *Cochrane Database Syst Rev* 2006;3:CD006066.
- [32] Sibanda J, Beard RW. Influence on clinical practice of routine intrapartum fetal monitoring. *Br Med J* 1975;3:341–3.
- [33] Liston R, Sawchuck D, Young D, Society of Obstetrics and Gynaecologists of Canada, British Columbia Perinatal Health Program. Fetal Health Surveillance: antepartum and intrapartum consensus guideline. *J Obstet Gynaecol Can* 2007;29(9 Suppl 4):S3–56. Erratum in: *J Obstet Gynaecol Can* 2007;29(11):909.
- [34] Brown P, Brunnhuber K, Chalkidou K, Chalmers I, Clarke M, Fenton M, et al. How to formulate research recommendations. *Br Med J* 2006;333:804–6.
- [35] Pettiti DB, Teutsch SM, Barton MB, Sawaya GF, Ockene JK, DeWitt T. Update on the methods of the U.S. Preventive Services Task Force: insufficient evidence. *Ann Intern Med* 2009;150:199–205.
- [36] Hussain T, Michel G, Shiffman RN. The Yale Guideline Recommendation Corpus: a representative sample of the knowledge content of guidelines. *Int J Med Inform* 2009;78:354–63.
- [37] Montori V, Devereaux P, Straus S, Haynes B, Guyatt G. Decision making and the patient. In: Guyatt G, Rennie D, Meade M, Cook D, editors. *The users’ guides to the medical literature: a manual for evidence-based clinical practice*. 2nd ed. New York, NY: McGraw-Hill; 2008.