ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 383–394

Journal of Clinical Epidemiology

GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables

Gordon Guyatt[a,b,*], Andrew D. Oxman[c], Elie A. Akl[m], Regina Kunz[d], Gunn Vist[c], Jan Brozek[a], Susan Norris[e], Yngve Falck-Ytter[f], Paul Glasziou[g], Hans deBeer[h], Roman Jaeschke[b], David Rind[i], Joerg Meerpohl[j,k], Philipp Dahm[l], Holger J. Schünemann[a,b]

[a]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[b]*Department of Medicine, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[c]*Norwegian Knowledge Centre for the Health Services, St. Olavs plass, 0130 Oslo, Norway*
[d]*Academy of Swiss Insurance Medicine, University Hospital Basel, Basel, Switzerland*
[e]*Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239-3098, USA*
[f]*Division of Gastroenterology, Case and VA Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA*
[g]*Bond University, Gold Coast, Australia*
[h]*Dutch Association of Nursing-home Specialists, Mercatorlaan 1200, 3528 BL Utrecht, The Netherlands*
[i]*Harvard Medical School, UpToDate, Boston, MA, USA*
[j]*German Cochrane Center, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, 79104 Freiburg, Germany*
[k]*Department of Pediatric and Adolescent Medicine, Division of Pediatric Hematology and Oncology, University Medical Center Freiburg, 79106 Freiburg, Germany*
[l]*Department of Urology, University of Florida, College of Medicine, Gainesville, FL 3210, USA*
[m]*Department of Medicine, State University of New York at Buffalo, Buffalo, NY, USA*

Accepted 8 April 2010

Abstract

This article is the first of a series providing guidance for use of the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system of rating quality of evidence and grading strength of recommendations in systematic reviews, health technology assessments (HTAs), and clinical practice guidelines addressing alternative management options. The GRADE process begins with asking an explicit question, including specification of all important outcomes. After the evidence is collected and summarized, GRADE provides explicit criteria for rating the quality of evidence that include study design, risk of bias, imprecision, inconsistency, indirectness, and magnitude of effect.

Recommendations are characterized as strong or weak (alternative terms conditional or discretionary) according to the quality of the supporting evidence and the balance between desirable and undesirable consequences of the alternative management options. GRADE suggests summarizing evidence in succinct, transparent, and informative summary of findings tables that show the quality of evidence and the magnitude of relative and absolute effects for each important outcome and/or as evidence profiles that provide, in addition, detailed information about the reason for the quality of evidence rating.

Subsequent articles in this series will address GRADE's approach to formulating questions, assessing quality of evidence, and developing recommendations.

Keywords: GRADE; systematic reviews; clinical practice guidelines; health technology assessment; quality of evidence; strength of recommendations

1. Introduction

In this, the first of a series of articles describing the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach to rating quality of evidence and grading strength of recommendations, we will briefly summarize what GRADE is, provide an overview of the GRADE process of developing recommendations, and present the endpoint of the GRADE evidence summary: the evidence profile (EP) and the summary of findings

The Grading of Recommendations Assessment, Development, and Evaluation (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* website.

* Corresponding author. CLARITY Research Group, Department of Clinical Epidemiology & Biostatistics, Room 2C12, 1200 Main Street West Hamilton, Ontario L8N 3Z5, Canada. Tel.: +905-525-9140; fax: +905-524-3841.

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

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2010.04.026

Key Points

- Grading of Recommendations Assessment, Development, and Evaluation (GRADE) offers a transparent and structured process for developing and presenting summaries of evidence, including its quality, for systematic reviews and recommendations in health care.
- GRADE provides guideline developers with a comprehensive and transparent framework for carrying out the steps involved in developing recommendations.
- GRADE's use is appropriate and helpful irrespective of the quality of the evidence: whether high or very low.
- Although the GRADE system makes judgments about quality of evidence and strength of recommendations in a systematic and transparent manner, it does not eliminate the inevitable need for judgments.

(SoFs) table. We will provide our perspective on GRADE's limitations and present our plan for this series.

2. What is GRADE?

GRADE offers a system for rating quality of evidence in systematic reviews and guidelines and grading strength of recommendations in guidelines. The system is designed for reviews and guidelines that examine alternative management strategies or interventions, which may include no intervention or current best management. In developing GRADE, we have considered a wide range of clinical questions, including diagnosis, screening, prevention, and therapy. Most of the examples in this series are clinical examples. The GRADE system can, however, also be applied to public health and health systems questions.

GRADE is much more than a rating system. It offers a transparent and structured process for developing and presenting evidence summaries for systematic reviews and guidelines in health care and for carrying out the steps involved in developing recommendations. GRADE specifies an approach to framing questions, choosing outcomes of interest and rating their importance, evaluating the evidence, and incorporating evidence with considerations of values and preferences of patients and society to arrive at recommendations. Furthermore, it provides clinicians and patients with a guide to using those recommendations in clinical practice and policy makers with a guide to their use in health policy.

A common definition of guidelines refers to ''systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances'' [1]. This series will describe GRADE's comprehensive approach to guideline development and to other similar guidance documents.

The optimal application of the GRADE approach requires systematic reviews of the impact of alternative management approaches on all patient-important outcomes. In the future, as specialty societies (e.g., American College of Physicians), national guideline developers and HTA agencies (e.g., National Institute for Health and Clinical Excellence), publishers (e.g., BMJ), publications (e.g., UpToDate), and international organizations (e.g., World Health Organization, Cochrane Collaboration) pool resources, high-quality evidence summaries will become increasingly available. As a result, even guideline panels with limited resources charged with generating recommendations for local consumption will be able to use GRADE to produce high-quality guidelines [2].

3. Purpose of this series

This series of articles about GRADE is most useful for three groups: authors of systematic reviews, groups conducting HTAs, and guideline developers. GRADE suggests somewhat different approaches for rating the quality of evidence for systematic reviews and for guidelines. HTA practitioners, depending on their mandate, can decide which approach is more suitable for their goals.

The GRADE approach is applicable irrespective of whether the quality of the relevant evidence is high or very low. Thus, all those who contribute to systematic reviews and HTA, or who participate in guideline panels, are likely to find this series informative. Consumers—and critics—of reviews and guidelines who desire an in-depth understanding of the evidence and recommendations they are using will also find the series of interest.

The series will provide a ''how to'' guide through the process of producing systematic reviews and guidelines, using examples to illustrate the concepts. We will not start with a broad overview of GRADE but rather assume that readers are familiar with the basics. Those who are not familiar may want to begin by reading a brief summary of the approach [3]. Those who want to start with a more detailed overview should examine all the articles in a previously published series describing the GRADE approach [4–9]. Finally, a computer program (GRADEpro) [10] and associated help file [11] that facilitate the development of EPs and SoFs tables provide a complement to this series.

4. The GRADE process—defining the question and collecting evidence

Figure 1 presents a schematic view of GRADE's process for developing recommendations in which unshaded boxes describe steps in the process common to systematic reviews

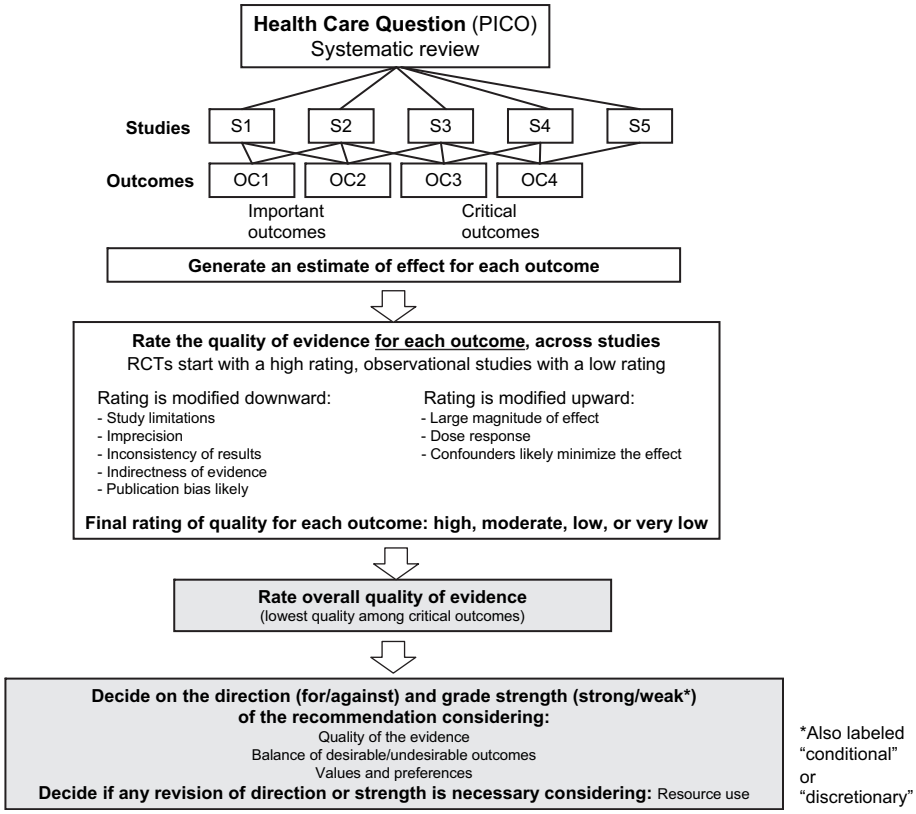

Fig. 1. Schematic view of GRADE’s process for developing recommendations. *Abbreviation*: RCT, randomized controlled trials.

and guidelines and the shaded boxes describe steps that are specific to guidelines. One begins by defining the question in terms of the populations, alternative management strategies (an intervention, sometimes experimental and a comparator, sometimes standard care), and all patient-important outcomes (in this case four) [12]. For guidelines, one classifies those outcomes as either critical (two outcomes in the figure) or important but not critical (two outcomes). A systematic search leads to inclusion of relevant studies (in this schematized presentation, five such studies).

Systematic review or guideline authors then use the data from the individual eligible studies to generate a best estimate of the effect on each patient-important outcome and an index (typically a confidence interval [CI]) of the uncertainty associated with that estimate.

5. The GRADE process—rating evidence quality

In the GRADE approach, randomized controlled trials (RCTs) start as high-quality evidence and observational studies as low-quality evidence supporting estimates of intervention effects. Five factors may lead to rating down the quality of evidence and three factors may lead to rating up (Fig. 2). Ultimately, the quality of evidence for each outcome falls into one of four categories from high to very low.

Systematic review and guideline authors use this approach to rate the quality of evidence for each outcome across studies (i.e., for a body of evidence). This does not mean rating each study as a single unit. Rather, GRADE is ‘‘outcome centric’’: rating is made for each outcome, and quality may differ—indeed, is likely to differ—from one outcome to another within a single study and across a body of evidence.

For example, in a series of unblinded RCTs measuring both the occurrence of stroke and all-cause mortality, it is possible that stroke—much more vulnerable to biased judgments—will be rated down for risk of bias, whereas all-cause mortality will not. Similarly, a series of studies in which very few patients are lost to follow-up for the outcome of death, and very many for the outcome of quality of life, is likely to result in judgments of lower quality for the latter outcome. Problems with indirectness may lead to rating down quality for one outcome and not another within a study or studies if, for example, fracture rates are measured using a surrogate (e.g., bone mineral density) but side effects are measured directly.

6. The GRADE process—grading recommendations

Guideline developers (but not systematic reviewers) then review all the information to make a final decision about which outcomes are critical and which are important and come to a final decision regarding the rating of overall quality of evidence.

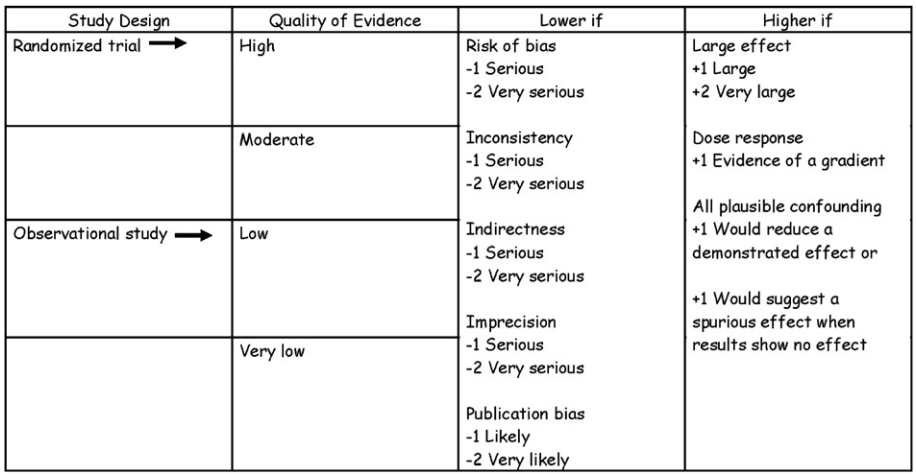

Fig. 2. Quality assessment criteria.

Guideline (but not systematic review) authors then consider the direction and strength of recommendation. The balance between desirable and undesirable outcomes and the application of patients' values and preferences determine the direction of the recommendation and these factors, along with the quality of the evidence, determine the strength of the recommendation. Both direction and strength may be modified after taking into account the resource use implications of the alternative management strategies.

7. The endpoint of the GRADE process

The endpoint for systematic reviews and for HTA restricted to evidence reports is a summary of the evidence—the quality rating for each outcome and the estimate of effect. For guideline developers and HTA that provide advice to policymakers, a summary of the evidence represents a key milestone on the path to a recommendation.

The GRADE working group has developed specific approaches to presenting the quality of the available evidence, the judgments that bear on the quality rating, and the effects of alternative management strategies on the outcomes of interest. We will now summarize these approaches, which we call the GRADE EP and the SoFs table. In doing so, we are taking something of a ''flashback'' approach to this series of articles: we begin by presenting the conclusion of the evidence summary process and will then go back to describe in detail the steps that are required to arrive at that conclusion.

8. What is the difference between an EP and a SoFs table?

An EP (Table 1) includes a detailed quality assessment in addition to a SoFs. That is, the EP includes an explicit judgment of each factor that determines the quality of evidence for each outcome (Fig. 2), in addition to a SoFs for each outcome. The SoF table (Table 2) includes an assessment of the quality of evidence for each outcome but not the detailed judgments on which that assessment is based.

The EP and the SoF table serve different purposes and are intended for different audiences. The EP provides a record of the judgments that were made by review or guideline authors. It is intended for review authors, those preparing SoF tables and anyone who questions a quality assessment. It helps those preparing SoF tables to ensure that the judgments they make are systematic and transparent and it allows others to inspect those judgments. Guideline panels should use EPs to ensure that they agree about the judgments underlying the quality assessments and to establish the judgments recorded in the SoF tables.

SoF tables are intended for a broader audience, including end users of systematic reviews and guidelines. They provide a concise summary of the key information that is needed by someone making a decision and, in the context of a guideline, provide a summary of the key information underlying a recommendation. GRADEpro computer software facilitates the process of developing both EPs and SoFs tables [10].

9. More than one systematic review may be needed for a single recommendation

Figure 1 illustrates that evidence must be summarized—the summaries ideally coming from optimally conducted systematic reviews—for each patient-important outcome. For each comparison of alternative management strategies, all outcomes should be presented together in one EP or SoFs table. It is likely that all studies relevant to a health care question will not provide evidence regarding every outcome. Figure 1, for example, shows the first study providing evidence for the first and second outcome, the

Table 1
GRADE evidence profile: antibiotics for children with acute otitis media

Quality assessment						Summary of findings					
						Number of patients			Absolute risk		
No of studies (Design)	Limitations	Inconsistency	Indirectness	Imprecision	Publication bias	Placebo	Antibiotics	Relative risk (95% CI)	Control risk[a]	Risk difference (95% CI)	Quality
Pain at 24h											
5 (RCT)	No serious limitations	No serious inconsistency	No serious indirectness	No serious imprecision	Undetected	241/605	223/624	RR 0.9 (0.78−1.04)	367/1,000	Not Significant	⊕⊕⊕⊕ High
Pain at 2−7 d											
10 (RCT)	No serious limitations	No serious inconsistency	No serious indirectness	No serious imprecision	Undetected	303/1,366	228/1,425	RR 0.72 (0.62−0.83)	257/1,000	72 fewer per 1,000 (44−98)	⊕⊕⊕⊕ High
Hearing, inferred from the surrogate outcome abnormal tympanometry—1 mo											
4 (RCT)	No serious limitations	No serious inconsistency	Serious indirectness (because of indirectness of outcome)	No serious imprecision	Undetected	168/460	153/467	RR 0.89 (0.75−1.07)	350/1,000	Not Significant	⊕⊕⊕○ Moderate
Hearing, inferred from the surrogate outcome abnormal tympanometry—3 mo											
3 (RCT)	No serious limitations	No serious inconsistency	Serious indirectness (because of indirectness of outcome)	No serious imprecision	Undetected	96/398	96/410	RR 0.97 (0.76−1.24)	234/1,000	Not Significant	⊕⊕⊕○ Moderate
Vomiting, diarrhea, or rash											
5 (RCT)	No serious limitations	Serious inconsistency (because of inconsistency in absolute effects)	No serious indirectness	No serious imprecision	Undetected	83/711	110/690	RR 1.38 (1.09−1.76)	113/1,000	43 more per 1,000 (10−86)	⊕⊕⊕○ Moderate

Abbreviations: GRADE, Grading of Recommendations Assessment, Development, and Evaluation; RCT, randomized controlled trials; CI, confidence interval; RR, risk ratio.
[a] The control rate is based on the median control group risk across studies.

Table 2
Summary of finding: antibiotics for acute otitis media in children

Antibiotics compared with placebo for acute otitis media in children

Patient or population: Children with acute otitis media
Setting: High- and middle-income countries
Intervention: Antibiotics
Comparison: Placebo

	Estimated risks (95% CI)					
	Control risk[a]	Intervention risk				
Outcomes	Placebo	Antibiotics	Relative effect (95% CI)	No. of Participants (studies)	Quality of the evidence (GRADE)	Comments
Pain at 24h	367 per 1,000	330 per 1,000 (286–382)	RR 0.9 (0.78–1.04)	1229 (5)	⊕⊕⊕⊕ High	
Pain at 2–7 d	257 per 1,000	185 per 1,000 (159–213)	RR 0.72 (0.62–0.83)	2791 (10)	⊕⊕⊕⊕ High	
Hearing, inferred from the surrogate outcome abnormal tympanometry—1 mo	350 per 1,000	311 per 1,000 (262–375)	RR 0.89 (0.75–1.07)	927 (4)	⊕⊕⊕◯ Moderate[b]	
Hearing, inferred from the surrogate outcome abnormal tympanometry—3 mo	234 per 1,000	227 per 1,000 (178–290)	RR 0.97 (0.76–1.24)	808 (3)	⊕⊕⊕◯ Moderate[b]	
Vomiting, diarrhea, or rash	113 per 1,000	156 per 1,000 (123–199)	RR 1.38 (1.09–1.76)	1,401 (5)	⊕⊕⊕◯ Moderate[c]	Ideally, evidence from nonotitis trials with similar ages and doses (not obtained) might improve the quality of the evidence.

Abbreviations: CI, confidence interval; RR, risk ratio; GRADE, Grading of Recommendations Assessment, Development, and Evaluation.

[a] The basis for the control risk is the median control group risk across studies. The intervention risk (and its 95% CI) is based on the control risk in the comparison group and the relative effect of the intervention (and its 95% CI).

[b] Because of indirectness of outcome.

[c] Generally, GRADE rates down for inconsistency in relative effects (which are not inconsistent in this case). Inconsistency here is in absolute effects, which range from 1% to 56%. Contributing factors to the decision to rate down in quality include the likely variation between antibiotics and the fact that most of the adverse events come from a single study. Consideration of indirect evidence from other trials of antibiotics in children (not undertaken) would likely further inform this issue.

second study for the first three outcomes, and so on. Indeed, there may be no overlap between studies providing evidence for one outcome and those providing evidence for another. For instance, RCTs may provide the relevant evidence for benefits and observational studies for rare, serious adverse effects.

Because most existing systematic reviews do not adequately address all relevant outcomes (many, for instance, are restricted to RCTs), the GRADE process may require relying on more than one systematic review. Ideally, future systematic reviews will comprehensively summarize evidence on all important outcomes for a relevant question.

10. A single systematic review may need more than one SoFs table

Systematic reviews often address more than one comparison. They may evaluate an intervention in two disparate populations or examine the effects of a number of interventions. Such reviews are likely to require more than one SoFs table. For example, a review of influenza vaccines may evaluate the effectiveness of vaccination for different populations, such as community dwelling and institutionalized elderly patients or for different types of vaccines.

11. An example of an EP

Table 1 presents an example of a GRADE EP addressing the desirable and undesirable consequences of use of antibiotics for children with otitis media living in high- and middle-income countries. The most difficult judgment in this table relates to the quality of evidence regarding adverse effects of antibiotics. In relative terms, the increases in adverse effects were reasonably consistent across trials. The trials, however, had very different rates of adverse effects (from 1% to 56%). Furthermore, from evidence external to the trials, we know that adverse effects differ across drugs (amoxicillin causes more adverse effects than penicillin). In addition, most of the events driving the increase come from a single trial which, of those included, had the highest risk of bias. The investigators recognized that ideally they would generate a summary of adverse effects from nonotitis trials with similar drug doses and patient age. Ultimately, they chose to rate down quality from high (starting high because the evidence comes from randomized trials) to moderate quality on the basis of inconsistency in absolute effects.

This dilemma faced by the investigators in making their rating of quality of evidence for adverse effects highlights two themes that will recur throughout this series. First, for many close-call judgments that are required in evaluating evidence, disagreement between reasonable individuals will be common. GRADE allows the pinpointing of the nature of the disagreement. Decision makers are then in a position to make their own judgments about the relevant issues.

Second, GRADE asks systematic review authors and guideline developers to consider quality of evidence under a number of discrete categories and to either rate down or not on the basis of each category (Fig. 2). Rigid adherence to this approach, however, ignores the fact that quality is actually a continuum and that an accumulation of limitations across categories can ultimately provide the impetus for rating down in quality. Ultimately, GRADE asks authors who decide to rate down quality by a single level to specify the one category most responsible for their decision (in this case, inconsistency of absolute effects) while documenting (as in the previous paragraph and in the footnotes in Tables 1 and 2), all factors that contributed to the final decision to rate down quality.

This presentation and the EP (Table 1) and SoF table (Table 2) illustrate another point: although we suggest standard formats based on pilot testing, user testing, and evaluations [13–16], alternative formats may be desirable for different audiences. Indeed, the order of the columns and the presentation of the absolute risks differs in the EP and SoF we present in this article.

In subsequent articles, we will continue to present examples of different formats for these tables. For both EPs and SoF tables, there is a trade-off between consistency, which facilitates their use and adaptation to address specific audiences or characteristics of the evidence, for example, by leaving out columns for some elements of the quality assessment or presenting the findings in a different way. Furthermore, EPs and SoF tables focusing on continuous variables and those addressing diagnostic questions may require a different format. Finally, the user testing conducted thus far is limited, and further testing may generate differing findings.

We suggest, however, that a few items should be included in all evidence summaries. For example, all EPs should include a row for each patient-important outcome. Typically, each row should include columns for the number of studies and the number of participants, the study design (randomized trials or observational studies), relevant factors that determine the evidence quality (Fig. 2), the overall judgment of quality (high, moderate, low, or very low) for that outcome, and estimates for the relative and absolute effects of the intervention.

12. An example of a SoFs table

Table 2 presents a SoF table in the format we recommend on the basis of pilot testing, user testing, and evaluations [10,12,13]. The Appendix presents an explanation of the terms found in the SoF table and the EP.

A SoF table presents the same information as the full EP, omitting the details of the quality assessment and adding a column for comments. The logic of the order of the columns is their importance—more important in the first columns and less important in the later. Aside from

Table 3
Examples of best practice statements and statements that could be confused with motherhood statements

Recommendations that are not helpful	Explanation	Recommendations that may be helpful but do not need grading	Explanation	Recommendations that need grading	Explanation
In patients presenting with chronic heart failure, take a careful and detailed history and perform a clinical examination.	"Careful and detailed history" is neither specific nor actionable.	In patients presenting with heart failure, initial assessment should be made of the patient's ability to perform routine/desired activities of daily living (LOE: C).	The alternative: initial assessment excluding ascertainment of ability to perform routine activities is not credible.	In patients with hypertension, the PE should include auscultation for carotid, abdominal, and femoral bruits.	This recommendation is specific, but may be a waste of time, or lead to positive results that lead to fruitless, resource-consuming investigation.
In patients with hypertension, the PE should include an appropriate measurement of BP, with verification in the contralateral arm.	It is not clear what exactly the authors mean by "appropriate measurement of BP."	Pregnant women should be offered evidence-based information and support to enable them to make informed decisions regarding their care, including details of where they will be seen and who will undertake their care (LOE: C).	Most would consider a recommendation to not offer such information a violation of basic standards of care.	In patients with diabetes, monofilaments should not be used to test more than 10 patients in one session and should be left for at least 24h to "recover" (buckling strength) between sessions (LOE: C).	If there is only very low-quality evidence to support such a recommendation, clinicians should be aware of this, and the recommendation should be weak.
All patients should undergo PE to define the severity of the hospital-acquired pneumonia, to exclude other potential sources of infection and to reveal specific conditions that can influence the likely etiologic pathogens (level II).	The elements of a PE that are necessary to reveal conditions that can influence the likely pathogens is uncertain.	Routinely record the daytime activities of people with schizophrenia in their care plans, including occupational outcomes.	A recommendation to omit recording such activities is not credible.	Monitoring for the development of diabetes in those with prediabetes should be performed every year (LOE: E).	The alternative should be specified (is it more frequently, less frequently, or not at all?). Specifying the alternative would make it evident that formal grading is desirable.
In patients presenting with a seizure, a PE (including cardiac, neurological, and mental state) and developmental assessment, where appropriate, should be carried out (LOE: C).	It is unclear what makes the particular aspects of PE or developmental assessment appropriate.	When working with caregivers of people with schizophrenia: provide written/verbal information on schizophrenia and its management, including how families/ caregivers can help through all phases of treatment.	Although randomized trials of specific educational programs may be warranted, a trial in which the basic information described here is withheld would be unacceptable.	Perform the A1C test at least two times a year in patients who are meeting treatment goals (and who have stable glycemic control) (LOE: E).	The alternative should be specified (is it more frequently, less frequently, or not at all?). Specifying the alternative would make it evident that formal grading is desirable.
Health care professionals should facilitate access as soon as possible to assessment/ treatment and promote early access throughout all phases of care.	The specific actions required to facilitate access are not specified and thus obscure.				

Abbreviations: BP, blood pressure; PE, physical examination; LOE, level of evidence.

a different order of columns, the SoF table (Table 2) presents the absolute risks in intervention and control groups with a CI around the intervention group rate, while the EP (Table 1) presents the risk difference with an associated CI. In addition, for nonsignificant outcomes (e.g., hearing, inferred from the surrogate outcome tympanometry) for the absolute risk difference, the EP notes only that results are nonsignificant, whereas the SoF table provides a CI around the intervention event rate.

The suggested format for SoF tables represents a compromise between simplicity (to make the information as easily accessible as possible to a wide audience) and completeness (to make the information and the underlying judgments as transparent as possible). When this format is used, judgments must still be made about what information to present (e.g., which outcomes and what levels of risk) and how to present that information (e.g., how to present continuous outcomes). As we have noted, although we encourage the use of this or a similar format and consistency, those preparing SoF tables should consider their target audience and the specific characteristics of the underlying evidence when deciding on the optimal format for a SoF table. Future editions of GRADEpro will include additional options for the preparation of EPs and SoF tables reflecting this flexibility [10].

13. Modifications of GRADE

Some organizations have used modified versions of the GRADE approach. We recommend against such modifications because the elements of the GRADE process are interlinked because modifications may confuse some users of evidence summaries and guidelines, and because such changes compromise the goal of a single system with which clinicians, policy makers, and patients can become familiar.

14. GRADE's Limitations

Those who want to use GRADE should consider five important limitations of the GRADE system. First, as noted previously, GRADE has been developed to address questions about alternative management strategies, interventions, or policies. It has not been developed for questions about risk or prognosis, although evidence regarding risk or prognosis may be relevant to estimating the magnitude of intervention effects or providing indirect evidence linking surrogate to patient-important outcomes.

Second, attempted application of GRADE to an ill-defined set of recommendations that one may call ''motherhood statements'' or ''good practice recommendations'' will prove problematic. A guideline panel may want to issue such recommendations relating to interventions that represent necessary and standard procedures of the clinical encounter or health care system—such as history taking and physical examination, helping patients to make informed decisions, obtaining written consent, or the importance of good communication. Some of these recommendations may not be helpful, and when they are helpful, it may not be a useful exercise to rate the quality of evidence or grade the strength of the recommendations. Other recommendations may be confused with good practice recommendations but may in fact require grading.

Recommendations that are unhelpful include those that are too vague to be implemented (e.g., ''take a comprehensive history'' or ''complete a detailed physical examination''). Some interpretations of such recommendations might lead to inefficient or counterproductive behavior. Guideline panels should issue recommendations only when they are both specific and actionable.

Recommendations that may be helpful but do not need grading are typically those in which it is sufficiently obvious that desirable effects outweigh undesirable effects that no direct evidence is available because no one would be foolish enough to conduct a study addressing the implicit clinical question. Typically, such recommendations are supported by a great deal of indirect evidence, but teasing out the nature of the indirect evidence would be challenging and a waste of time and energy. One way of recognizing such questions is that if one made the alternative explicit, it would be bizarre or laughable.

Procedures may be sufficiently ingrained in standard clinical practice that guideline panels would be inclined to consider them good practice recommendations when in fact a dispassionate consideration would suggest that legitimate doubt remains regarding the balance of desirable and undesirable consequences. Such recommendations should undergo formal rating of quality of evidence and grading of strength of recommendations. Table 3 provides examples of unhelpful good practice recommendations, helpful good practice recommendations, and recommendations that might be confused with good practice recommendations but require rating of quality of evidence and grading of recommendations.

Third, as illustrated in Fig. 3, preparing a guideline entails several steps both before and after those steps to which the GRADE system applies. It is important for review authors and guideline developers to understand where GRADE fits into the overall process and to look elsewhere for guidance related to those other steps [17,18]. We do, however, in later articles in this series, provide our view of how the GRADE system is best implemented in the context of these other steps.

Fourth, the overwhelming experience with GRADE thus far is in evaluation of preventive and therapeutic interventions and in addressing clinical questions rather than public health and health systems questions. Those applying GRADE to questions about diagnostic tests, to public health, or to health systems questions will face some special challenges [8,19]. We will address these challenges, particularly those related to diagnostic tests, later in this

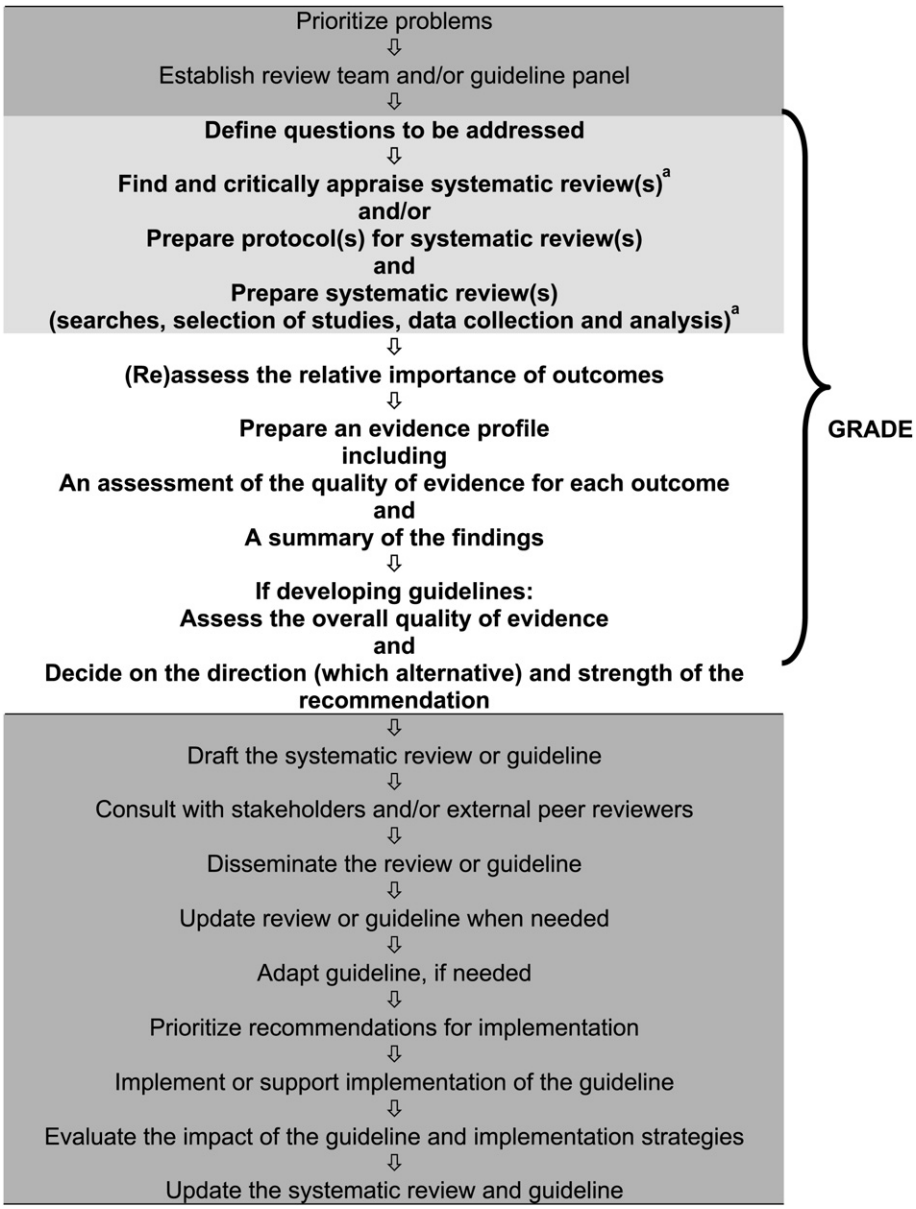

Fig. 3. Where GRADE fits in to the guideline development process. *Abbreviation*: GRADE, Grading of Recommendations Assessment, Development, and Evaluation. [a]Some aspects of the development and appraisal of systematic reviews fall clearly within the GRADE process and others do not. Particularly relevant to GRADE are the definition of the question and most particularly the definition of the outcomes, including the choice of the most important; the specification of a priori hypotheses to explain possible heterogeneity; and the interpretation of the results, in particular the generation of estimates of absolute effect and the interpretation of subgroup analyses.

series. Aware that work remains to be done in refining the GRADE process and addressing areas of uncertainty, the GRADE working group continues to meet regularly and continues to welcome new members to participate in the discussions.

Finally, GRADE will disappoint those who hope for a framework that eliminates disagreements in interpreting evidence and in deciding on the best among alternative courses of action. Although the GRADE system makes judgments about quality of evidence and strength of recommendations in a more systematic and transparent manner, it does not eliminate the need for judgments.

15. Where from here

The next article in this series will describe GRADE's approach to framing the question that a systematic review or guideline is addressing and deciding on the importance of outcomes. The next set of articles in the series will address in detail the decisions required to generate EPs and SoF tables, such as those presented in Tables 1 and 2. The series will then address special challenges related to diagnostic tests and resource use and the process of going from evidence to recommendations. The series will conclude by commenting on issues of applying GRADE in guideline panels.

Appendix. Explanations for SoFs tables (Table 2) and EPs (Table 1)

Examples from table	Explanations
Outcomes	Outcomes The tables provide the findings for the most important outcomes for someone making a decision. These include potential benefits and harms, whether the included studies provide data for these outcomes or not. Additional findings may be reported elsewhere in the review.
Absolute risks	Absolute risks Risk is the probability of an outcome occurring. The estimated risks columns in the SoF table present the best estimate of the risk in the control group (control risk in the EP) and the risk in the intervention group (intervention risk antibiotics), with a CI around the risk in the intervention group. If one wants to know the difference in absolute risk or the CI around the difference in risk, this requires subtraction. In the EP, the risk difference is presented directly.
185 per 1,000 (159−213)	Confidence interval A CI is a range around an estimate that conveys how precise the estimate is; in this example, the result is the estimate of the intervention risk (see in the following). The CI is a guide to how sure we can be about the quantity we are interested in (here the true absolute effect). The narrower the range between the two numbers, the more confident we can be about what the true value is; the wider the range, the less sure we can be. The width of the CI reflects the extent to which chance may be responsible for the observed estimate (with a wider interval reflecting more chance).
(95% CI)	95% CI As explained previously, the CI indicates the extent to which chance may be responsible for the observed numbers. In the simplest terms, a 95% CI means that we can be 95% confident that the true size of effect is between the lower and upper confidence limit (e.g., 0.62 and 0.83 in the example of a relative effect of pain at 2−7 d in Table 2). Conversely, there is a 5% chance that the true effect is outside of this range.
Estimated risk control 257 per 1,000	Estimated control risk (without the intervention) Estimated risks control (control rate in the EP) are typical rates of an outcome occurring without the intervention. They will ideally be based on studies of incidence in representative populations. Alternatively, if such observational studies are not available, they can be based on control group risks in comparative studies. When only one control group risk is provided, it is normally the median control group risk across the studies that provided data for that outcome. In this example (pain at 2−7 d), the risk of 257 events occurring in every 1,000 people indicates what would happen in a typical control group population. When relevant, the tables will provide information for more than one population, for instance differentiating between people at low and high risk when there are potentially important differences.
Intervention risk antibiotics 185 per 1,000 (159−213)	Intervention risk In this example, the estimated risk in the control group was 257 events in every 1,000 persons. Implementing the intervention in this population would result in a intervention intervention group risk of 185 events in every 1,000 people, given the pooled risk ratio (RR) across studies. If the table provides more than one control risk for an outcome, for instance differentiating between people at low and high risk, then a intervention risk is provided for each population. Determining the effect of the intervention requires subtraction. In the EP, the subtraction has been done for you. The intervention results in 72 fewer children in every 1,000 experiencing pain at 2−7 d.
Relative effect (95% CI)	Relative effect or RR
RR 0.72 (0.62−0.83)	Relative effects are ratios. Here the relative effect is expressed as a RR. Risk is the probability of an outcome occurring. A RR is the *ratio* between the risk in the intervention group and the risk in the control group. If the risk in the intervention group is 1% (10 per 1,000) and the risk in the control group is 10% (100 per 1,000), the relative effect is 10/100 or 0.10. If the RR is exactly 1.0, this means that there is no difference between the occurrence of the outcome in the intervention and the control group. It is unusual for the RR to be exactly 1.0, and what it means if it is above or below this value depends on whether the outcome being counted is judged to be good or bad. If the RR is greater than 1.0, the intervention increases the risk of the outcome. If it is a good outcome (for example, the birth of a healthy baby), an RR greater than 1.0 indicates a desirable effect for the intervention; whereas, if the outcome is bad (for example, death), an RR greater than 1.0 would indicate an undesirable effect. If the RR is less than 1.0, the intervention decreases the risk of the outcome. This indicates a desirable effect, if it is a bad outcome (for example, death) and an undesirable effect if it is a good outcome (for example, birth of a healthy baby).
The mean edema score in the intervention groups was on average 4.7 lower (95% CI −4.5, −4.9).	There are no mean scores in this example (but this is what it would look like if there were).

(*Continued*)

Appendix. Continued

Examples from table	Explanations
2,791 (10 studies)	Number of participants (studies) The table provides the total number (no.) of participants across studies (2,791 in this example) and the number of studies (10) that provided data for that outcome. This indicates how much evidence there is for the outcome. The EP includes columns that provide the number of events and number of patients, in each of the control (241/1,605) and intervention (223/1,624) groups
Quality of the evidence (GRADE)	Quality of the evidence The quality of the evidence is a judgment about the extent to which we can be confident that the estimates of effect are correct. These judgments are made using the GRADE system and are provided for each outcome. The judgments are based on the type of study design (randomized trials vs. observational studies), the risk of bias, the consistency of the results across studies, and the precision of the overall estimate across studies. For each outcome, the quality of the evidence is rated as high, moderate, low, or very low.
	A blank space indicates that the information is not relevant.
	What is the difference between the risks presented in the shaded columns and the relative effect? The effect of an intervention can be described by comparing the risk of the control group with the risk of the intervention group. Such a comparison can be made in different ways.
	One way to compare two risks is to calculate the *difference* between the risks. This is the absolute effect. The absolute effect can be found in the SOFs table by calculating the difference between the numbers in the shaded columns—the control risk in the control group on the left and the intervention risk in the intervention group on the right. The EP does the subtraction for you.
	Here is an example: Consider the risk for blindness in a patient with diabetes over a 5-year period. If the risk for blindness is found to be 20 in 1,000 (2%) in a group of patients treated conventionally and 10 in 1,000 (1%) in patients treated with a new drug, the absolute effect is derived by subtracting the intervention group risk from the control group risk: 2%−1% = 1%. Expressed in this way, it can be said that the new drug reduces the 5-year risk for blindness by 1% (absolute effect is 10 fewer per 1,000).
	Another way to compare risks is to calculate the *ratio* of the two risks. Given the data above, the relative effect is derived by *dividing* the two risks, with the intervention risk being divided by the control risk: 1% ÷ 2% = ½ (0.50). Expressed in this way, as the ''relative effect,'' the 5-year risk for blindness with the new drug is one-half the risk with the conventional drug.
	Here the table presents risks as times per 1,000 instead of percentage, as this tends to be easier to understand. Whenever possible, the table presents the relative effect as the RR.
	Usually the absolute effect is different for groups that are at high and low risk, whereas the relative effect often is the same. Therefore, when it is relevant, GRADE tables report risks for groups at different levels of risk.

References

[1] Field M, Lohr K. Clinical practice guidelines: directions for a new program. Washington, DC: National Academic Press; 1990.

[2] 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.

[3] Guyatt G, Vist G, Falck-Ytter Y, Kunz R, Magrini N, Schunemann H. An emerging consensus on grading recommendations? ACP J Club 2006;144(1):A8−9.

[4] Guyatt GH, Oxman AD, Kunz R, Jaeschke R, Helfand M, Liberati A, et al. Incorporating considerations of resources use into grading recommendations. BMJ 2008;336:1170−3.

[5] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924−6.

[6] Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is ''quality of evidence'' and why is it important to clinicians? BMJ 2008;336:995−8.

[7] Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al. Going from evidence to recommendations. BMJ 2008;336:1049−51.

[8] Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106−10.

[9] Jaeschke R, Guyatt GH, Dellinger P, Schunemann H, Levy MM, Kunz R, et al. Use of GRADE grid to reach decisions on clinical practice guidelines when consensus is elusive. BMJ 2008;337:a744.

[10] Brozek J, Oxman A, Schünemann HJ. GRADEpro. [Computer program]. Version 3.2 for Windows. Available at http://mcmaster.flintbox.com/technology.asp?Page=3993 and http://www.cc-ims.net/revman/gradepro. Accessed October 21, 2010.

[11] Schünemann H, Brozek J, Guyatt G, Oxman A. GRADE handbook for grading quality of evidence and strength of recommendation; 2010.

[12] Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group. JAMA 1993;270:2093−5.

[13] 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.

[14] GRADE Working Group. Systems for grading the quality of evidence and the strength of recommendations II: Pilot study of a new system. BMC Health Serv Res 2005;5:25.

[15] Rosenbaum SE, Glenton C, Nylund HK, Oxman AD. User testing and stakeholder feedback contributed to the development of understandable and useful Summary of Findings tables for Cochrane reviews. J Clin Epidemiol 2010;63:607−19.

[16] Rosenbaum S, Glenton C, Oxman A. Evaluation of summary of findings tables for Cochrane reviews. J Clin Epidemiol 2010;63:620−6.

[17] Schunemann H, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 1. Guidelines for guidelines. Health Res Policy Syst 2006;4:13.

[18] Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0. [www.cochrane-handbook.org].

[19] Schünemann H, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 9. Grading evidence and recommendations. Health Res Policy Syst 2006;4:21.