

## Editorial

# Formulating Recommendations With GRADE: A Matter of Confidence

## La formulación de recomendaciones con GRADE: cuestión de confianza

Pablo Alonso-Coello,<sup>a,b,\*</sup> Ivan Solà,<sup>a,b</sup> and Ignacio Ferreira-González<sup>b,c</sup>

<sup>a</sup> Centro Cochrane Iberoamericano, Instituto de Investigación Biomédica Sant Pau (IIB-Sant Pau), Barcelona, Spain

<sup>b</sup> CIBER de Epidemiología Clínica y Salud Pública (CIBERESP), Barcelona, Spain

<sup>c</sup> Unidad de Epidemiología, Servicio de Cardiología, Hospital Vall d'Hebron, Barcelona, Spain

Article history:

Available online 23 November 2012

### CLINICAL PRACTICE GUIDELINES: THE CURRENT SITUATION

Clinical practice guidelines (CPG), understood to be a series of recommendations to guide clinical practice, have become very popular tools among health care professionals. Their development has taken place both on the international level and in the Spanish medical setting, and has been accompanied by considerable advances in the methodology for developing and evaluating them.<sup>1,2</sup>

The development of CPG has incorporated such important changes as the multidisciplinary composition of the groups that prepare them, the obligation of group members to declare conflicts of interest, systematic reviews of the literature, and a more precise and structured formulation of the recommendations. However, this development of the methodology has not always resulted in higher quality.<sup>3</sup>

Specifically, in the case of the strategies for formulating recommendations, there has been an excessive proliferation of approaches that has led to a multiplicity of systems. This fact, as well as a number of limitations, has complicated communication among those designing the CPG and has often confused end users. Thus, an international effort has recently been proposed in the attempt to reach a consensus on a single system that overcomes the aforementioned limitations.<sup>4,5</sup> The proposal, known as GRADE (*Grading of Recommendations Assessment, Development and Evaluation*), was drawn up by an international group of CPG developers, clinicians, and methodologists belonging to the major organizations involved in preparing CPG. This proposal is quickly being adopted by numerous institutions all over the world, such as the National Institute of Clinical Excellence (NICE), the World Health Organization (WHO), and the Cochrane Collaboration, and publications like Clinical Evidence or Uptodate ([www.gradeworkinggroup.org](http://www.gradeworkinggroup.org)).<sup>6</sup> In our context, the National Program for the Preparation of Clinical Practice Guidelines of the Spanish Health System ([www.guiasalud.es](http://www.guiasalud.es)),<sup>1</sup> among others, has begun to utilize it.

The most relevant differences between GRADE and other previous systems can be summarized as follows: a) grading of the importance of outcomes of interest (for example, acute myocardial

infarction); b) explicit separation between the quality of the evidence and the strength of the recommendations; c) use of explicit criteria for the evaluation of evidence quality and the strength of the recommendations; and d) consideration of patient values and preferences in the formulation of recommendations. The purpose of this article is to describe the GRADE system to CPG users. For more in-depth information, those interested can consult two series of articles, one published in the British Medical Journal<sup>4</sup> and the other, aimed mainly at CPG developers, in the Journal of Clinical Epidemiology.<sup>5</sup>

### THE NEED TO EVALUATE WHAT IS RELEVANT

When we have to make a decision as to whether an intervention entails greater benefits than risks, not all the outcomes of interest are equally important; thus, our judgment should be based on those most important for decision making. GRADE proposes a classification of the importance of the outcomes of interest according to 3 categories: key, important, and not important.<sup>7</sup> Key outcomes will be those that determine the quality of the evidence and, ultimately, the balance between the benefits and risks, and the strength of the recommendations. For example, in the recent CPG for antithrombotic therapy for patients with atrial fibrillation, the guideline developers chose mortality and nonfatal stroke as key outcomes.<sup>8</sup> Severe (nonfatal) extracranial bleeding and systemic embolism were considered important, but not key. Finally, GRADE promotes the consideration of the perspective of the patients in the evaluation of the importance of the outcomes since their values and preferences may not coincide with those of the guideline developers.

### WHAT DOES QUALITY OF EVIDENCE ENTAIL?

Users of CPG should have access to a simple method to determine the degree of confidence they can place in the results gathered from the review of the available literature. This information is crucial and, moreover, is highly relevant for grading the strength of the recommendations. The GRADE system defines quality of evidence as the degree to which our confidence in the estimate of a given effect (for example, reducing the risk of nonfatal stroke by 50%) is adequate to support a recommendation.<sup>7</sup> Evidence quality is evaluated for each of the key outcomes. Thus,

\* Corresponding author: Centro Cochrane Iberoamericano, Instituto de Investigación Biomédica Sant Pau (IIB-Sant Pau), Sant Antoni Maria Claret 171, 08041 Barcelona, Spain.

E-mail address: [palonso@santpau.cat](mailto:palonso@santpau.cat) (P. Alonso-Coello).

for an intervention of interest involving a given comparison (for example, dabigatran versus warfarin), there can be different classifications of the evidence quality. For a specific outcome (for example, nonfatal stroke), we may find a series of studies in which there are no limitations in the design, and for another relevant outcome (for example, death) the results reported may be less precise. Thus, our degree of confidence for each of these two outcomes will differ (high and moderate, respectively).

Likewise, the proposal for the classification of evidence quality according to GRADE not only takes into account the risk of bias, as do other systems, but considers other factors as well, for example, the consistency or precision of the results (Figure). GRADE proposes a classification with 4 categories (high, moderate, low, and very low). Both for randomized clinical trials (the quality of which is initially considered to be high) and for observational studies (the quality of which is initially considered to be low), different factors can reduce (or increase) our confidence in the estimate of the effect observed. These factors are: *a*) limitations in the design and implementation of the studies (risk of bias); *b*) heterogeneity of the results; *c*) the absence of direct evidence (understood to be the absence in the literature of proof directly applicable to the patients, interventions, or outcome of the clinical situation); *d*) imprecision of the results; and *e*) publication bias (Figure). The presence of one or more limitations related to these factors will decrease the quality by one level (for example, from high to moderate) or more. For instance, results of clinical trials evaluating beneficial changes in arterial blood pressure (indirect evidence concerning outcomes such as stroke) associated with dietary advice are heterogeneous. In this case, there are 2 limitations that would affect evidence quality: heterogeneity and indirect evidence. Thus, depending on the variability observed and the indirectness of the evidence, the quality could be classified as moderate or even low.

A few circumstances can lead to an increased confidence in the results of observational studies.<sup>7</sup> In these situations, we should only consider whether there is any reason to question the quality of the evidence due to limitations in the design or implementation. Two paradigmatic examples are the use of insulin in the treatment of diabetic ketoacidosis or of adrenalin in anaphylaxis. The lack of

randomized clinical trials does not impede us from having a high degree of confidence in their effectiveness. The reasons for the increased confidence are the presence of a considerable and immediate effect of the treatments and a radical change in the prognosis of these patients since the introduction of these therapies (Figure).

The GRADE system makes it possible to condense the available information in a structured summary of findings (SoF) table. This table includes the number of studies available for each key outcome of interest, the quality of the evidence, and the estimators of the observed effect, in relative and absolute terms, among other data. This table is intended for users of systematic reviews and of CPG, and can be created with GRADEpro software, which is available as a free download.

In the example of the CPG for antithrombotic therapy mentioned above,<sup>8</sup> the purpose was to evaluate the evidence available to formulate a recommendation concerning the use of dabigatran compared to warfarin (Table). In this case, the evidence quality was reduced in 3 of the outcomes of interest (mortality, nonfatal extracranial bleeding, and systemic thromboembolism) because of the imprecision observed in the results (the confidence intervals included both a potential benefit of dabigatran and the absence of effect, or even an increase in the risk of an undesirable outcome). According to the available information, dabigatran would prevent 3 strokes per year, compared to warfarin, and would probably reduce the risk of death in 1 of every 1000 patients treated per year. Moreover, it does not appear to increase the risk of severe extracranial bleeding or modify the risk of systemic embolism.

#### CAN WE TRUST THAT A RECOMMENDATION WILL PRODUCE MORE BENEFITS THAN RISKS?

GRADE defines the strength of a recommendation in terms of our confidence in the desired outcomes of an intervention (for example, its benefits) outweighing the undesirable outcomes (for example, inconveniences or adverse effects of a treatment).<sup>9</sup> The GRADE system divides recommendations into 4 categories

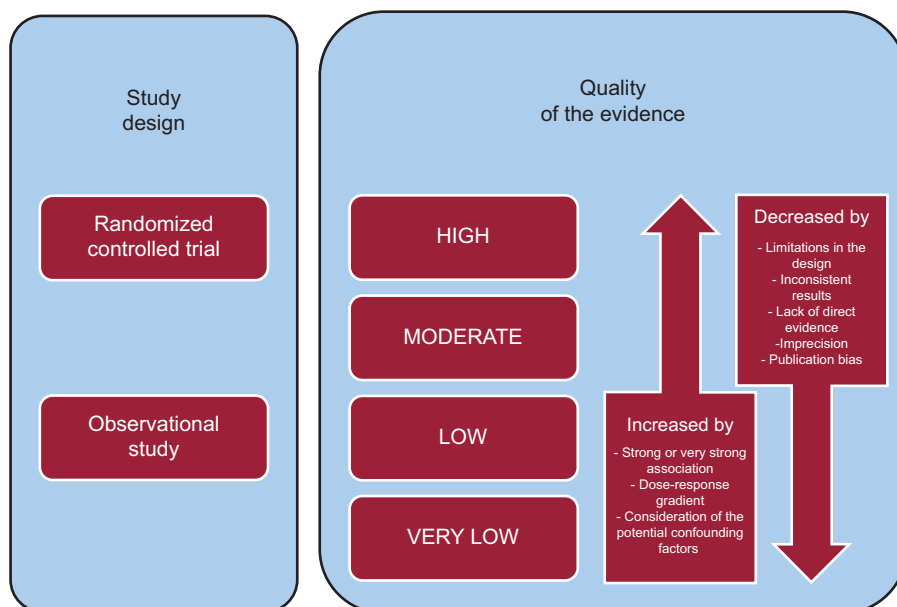







Figure. Quality of the evidence and factors that modify it.

**Table**Summary of the Findings. Comparison Between Dabigatran and Warfarin in Patients With Atrial Fibrillation and Intermediate Risk for Stroke<sup>8</sup> (CHADS<sub>2</sub>=1)

Outcomes	Participants (studies) Follow-up	Quality of the evidence (GRADE)	Relative effect (95%CI)	Estimated absolute effect at 1 year	
				Risk with warfarin	Risk with dabigatran (95% CI)
<i>Death</i>	12 098 (1) Mean 2 years	 Moderate due to imprecision <sup>a</sup>	RR=0.89 (0.79-1.01)	38 deaths per 1000 <sup>b</sup>	4 fewer deaths per 1000 (from 8 fewer to 0 more)
<i>Stroke</i> Includes ischemic stroke and nonfatal intracranial bleeding <sup>b</sup>	12 098 (1) Mean 2 years	 High	RR=0.67 (0.52-0.86)	Intermediate risk of stroke (CHADS <sub>2</sub> =1)	
				8 strokes per 1000	3 fewer strokes per 1000 (from 1 fewer to 4 fewer)
<i>Severe extracranial bleeding</i> Includes nonfatal major extracranial bleeding	12 098 (1) Mean 2 years	 Moderate due to imprecision <sup>a</sup>	RR=1.07 (0.91-1.26)	13 cases of bleeding per 1000	1 more case of bleeding per 1000 (from 1 fewer to 3 more)
<i>Systemic embolism</i>	12 098 (1) Mean 2 years	 Moderate due to imprecision <sup>a</sup>	RR=0.85 (0.39-1.84)	2 embolisms per 1000	0 fewer embolisms per 1000 (from 1 fewer to 2 more)
<i>Inconveniences of the treatment</i>	NA	 High		Warfarin: daily medication, lifestyle limitations, dietary restrictions, frequent blood tests and visits to the doctor Dabigatran: daily medication	

95%CI, 95% confidence interval; CHADS<sub>2</sub>, congestive heart failure, hypertension, age  $\geq 75$  years, diabetes mellitus, and stroke; GRADE, Grading of Recommendations Assessment, Development and Evaluation; NA, not available; RR, relative risk.

<sup>a</sup>The 95% confidence interval does not rule out the possibility of an appreciable risk or benefit with dabigatran therapy.

<sup>b</sup>Intracranial bleeding includes intracerebral, subdural, and subarachnoid bleeding.

depending on its formulation (either for or against) and the strength of the recommendation (strong or weak). In a recommendation in favor of a given option, the desired effects of one intervention versus another outweigh the undesirable effects. In a recommendation against an option, the undesirable effects of one intervention versus another outweigh the desired effects. In the case of a strong recommendation, we can trust there to be a favorable balance between the desired and the undesirable effects of one intervention versus another. In contrast, in a weak one, there is uncertainty with respect to this balance.

The implications of strong and weak recommendations for patients, health care professionals, and health care managers differ. For instance, in the case of the patients, a strong recommendation implies that the majority of them would agree with the recommended intervention and that only a small number would not. In contrast, a weak recommendation implies that most people would agree with the recommended action, but a considerable number of them would not. In the case of health care professionals, a strong recommendation would indicate that the majority of the patients should undergo the recommended intervention. In the case of a weak recommendation, different options could be appropriate, and the physician should help each patient reach a decision as much in accordance with his or her values and preferences as possible.

To determine the strength of a recommendation and whether it should be formulated in favor of or against an intervention, GRADE considers 4 factors:

- **Balance between benefits and risks.** When the difference between the desired and undesirable outcomes of the intervention is very great, the formulation of a strong recommendation (either for or against) is more likely. In contrast, when the difference is small, a weak recommendation is usually formulated. For example, whereas the risk-benefit balance of thrombolysis within the first 6 h of myocardial infarction is clearly tipped in favor of the benefits, from 6 h on the difference is not that important.
- **Quality of the evidence.** Before formulating a recommendation, it is necessary to know the confidence in the quality or the confidence in the estimate of the effects reported in the literature. If the quality of the evidence is low, the formulation of a weak recommendation is more likely. In contrast, if the quality is high, the formulation of a strong recommendation is more likely. However, there are situations in which a strong recommendation can be justified although the evidence quality is low or very low. For example:
  - When the quality of evidence is low compared to the benefits of an intervention in a life-threatening situation (strong recommendation in favor of some action), as for example in the case of emergency surgery for ventricular free wall rupture in acute myocardial infarction.
  - When the quality of the evidence is low compared to the benefits of an intervention is low, and high for potential damage or for the very high cost of the intervention (strong recommendation against). For example, although an implantable defibrillator could have certain potential benefits in

patients with an ejection fraction greater than 40% one month after myocardial infarction, the magnitude of the benefits may not justify the high cost.

- When the quality of the evidence indicating equivalence between two interventions is low, but is high for potentially lesser damage with one of the alternatives (strong recommendation for the intervention associated with fewer adverse events).
- When the quality of the evidence showing equivalence between two interventions is high, but is low for damage from one of the alternatives (strong recommendation for the intervention with fewer adverse events). One example would be use of acetylsalicylic acid (ASA) versus paracetamol in children with fever and measles. The quality of the evidence indicating their efficacy is similarly high, but is low for the association between ASA and Reye syndrome.
- **Values and preferences.** Weighing the benefits and risks of different therapeutic and diagnostic strategies inevitably requires making value judgments. Ideally, to carry out this process we should know the values and preferences of our patient population and to what extent they differ from one individual to the next. However, we often do not have this information or do not know to what extent the values and preferences are uniform; thus, in these cases the recommendations will probably be more prudent or weaker. For instance, again using the comparison between dabigatran and warfarin in atrial fibrillation,<sup>8</sup> we should place in the balance a reduction in the risk of stroke and an increase in the risk of extracranial bleeding. The available literature shows that, in general, patients consider the importance of preventing stroke as 3-fold greater than that of preventing extracranial bleeding. Nevertheless, the variability observed in the available studies would probably result in the formulation of strong recommendations only when the benefits were much greater than the risks or vice versa, or under circumstances in which the values and preferences were relatively uniform. In the case of the example regarding atrial fibrillation, and due to the fact that the potential comparisons of the interventions reveal no differences in mortality (per 1000 patients treated over 1 year), if the number of strokes prevented is lower than one third of the number of cases of severe extracranial bleeding caused, the recommendation is contrary to the application of the intervention of interest. In the case of the strokes prevented, if the number is substantially higher than one third of the cases of severe extracranial bleeding caused by the antithrombotic therapy being evaluated, the recommendations formulated are in favor of the intervention.
- **Costs and use of resources.** The costs, in contrast to other factors, are more difficult to evaluate due to the fact that there is considerable variability in the resources involved, the setting, and the timing. A high cost reduces the probability of formulating strong recommendations for a given intervention. Nevertheless, the context can prove to be critical when it comes to making the final decision.

## JUSTIFICATION OF THE RECOMMENDATION

The above 4 factors should be incorporated and weighed when determining the strength of the recommendations. For this purpose, it is essential that the groups developing CPG reflect this process explicitly in table form. In the case of the example involving dabigatran and warfarin, with respect to the balance between benefit and risk, dabigatran prevents 3 strokes for every 1000 patients with atrial fibrillation and moderate risk of stroke (CHADS=1), but produces 1 case of additional extracranial

bleeding. The evidence quality is moderate due to the imprecise results regarding the outcomes of death, extracranial bleeding, and systemic embolism. On the other hand, there is potential variability in patient values and preferences concerning treatment and dabigatran is costly, although it is probably cost-effective in patients at moderate or high risk of stroke. The group developing these guidelines also took into account other factors in this case, such as the lack of long-term data on safety and efficacy of dabigatran and the absence of an antidote. In fact, they mention that it would be reasonable to continue that approach in patients with well-controlled oral anticoagulation therapy, rather than change to dabigatran. The incorporation of these factors leads to the formulation of a weak recommendation in favor of dabigatran, and it is proposed that, in patients with atrial fibrillation and moderate risk of stroke, this agent be considered in place of warfarin (weak recommendation in favor).<sup>8</sup>

## CONCLUSIONS

GRADE is a rigorous system for the evaluation of the quality and formulation of recommendations that addresses the limitations of previous systems. It provides the groups developing guidelines with an explicit and structured framework, but does not eliminate the need to make judgments when it comes to the many decisions that must be made in drawing up recommendations. At the present time, a large number of institutions have begun to use the GRADE system, and its implementation and influence are increasingly widespread both in Spain and on the international level. In this respect, the European Society of Cardiology has begun to introduce it in some of its most recent guidelines<sup>10</sup> and, thus, there are reasons to believe that the cardiology guidelines in Europe and Spain could soon have a common system for the formulation of recommendations.

## CONFLICTS OF INTEREST

Pablo Alonso-Coello and Ivan Solà are members of the GRADE working group.

## REFERENCES

1. Grupo de trabajo sobre GPC. Elaboración de Guías de Práctica Clínica en el Sistema Nacional de Salud. Manual Metodológico. Madrid: Plan Nacional para el SNS del MSC. Instituto Aragonés de Ciencias de la Salud-I+CS; 2007. Guías de Práctica Clínica en el SNS: I+CS. N.º 2006/OI [cited 12 Jul 2012]. Available at: <http://portal.guiasalud.es/emanuales/elaboracion/index-02.html>
2. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. AGREE Next Steps Consortium. AGREE II: advancing guideline development, reporting and evaluation in health care. *CMAJ*. 2010;182:E839–42.
3. Alonso-Coello P, Irfan A, Sola I, Gich I, Delgado-Noguera M, Rigau D, et al. The quality of clinical practice guidelines over the last two decades: a systematic review of guideline appraisal studies. *Qual Safety Health Care*. 2010;19:e58.
4. Guyatt GH, Oxman AD, Vist G, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al.; for the GRADE Working Group. Rating quality of evidence and strength of recommendations GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336:924–6.
5. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: A new series of articles in the *Journal of Clinical Epidemiology*. *J Clin Epidemiol*. 2011;64:380–2.
6. Grupo de trabajo GRADE. Grading of Recommendations Assessment, Development and Evaluation [cited 17 Jul 2012]. Available at: [www.gradeworking-group.org](http://www.gradeworking-group.org)
7. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ; GRADE Working Group. Rating quality of evidence and strength of recommendations: What is “quality of evidence” and why is it important to clinicians? *BMJ*. 2008;336:995–8.
8. You JJ, Singer DE, Howard PA, Lane DA, Eckman MH, Fang MC, et al. Antithrombotic therapy for atrial fibrillation: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141 Suppl 2:e531S–75S.

9. Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al.; GRADE Working Group. Rating quality of evidence and strength of recommendations: Going from evidence to recommendations. *BMJ*. 2008;336:1049–51.
10. Perk J, De Backer G, Gohlke H, Graham I, Reiner Z, Verschuren M, et al. European Guidelines on cardiovascular disease prevention in clinical practice (version 2012): The Fifth Joint Task Force of the European Society of Cardiology and

Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of nine societies and by invited experts) \*Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *Eur Heart J*. 2012; 33:1635-701 [cited 12 Jul 2012]. Available at: <http://www.escardio.org/guidelines-surveys/esc-guidelines/GuidelinesDocuments/guidelines-CVD-prevention.pdf>