ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 151−157

Journal of Clinical Epidemiology

GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes

Gordon Guyatt[a,j,*], Andrew D. Oxman[b], Shahnaz Sultan[c], Jan Brozek[d], Paul Glasziou[e], Pablo Alonso-Coello[f], David Atkins[g], Regina Kunz[h,i], Victor Montori[j], Roman Jaeschke[k], David Rind[l], Philipp Dahm[m], Elie A. Akl[n], Joerg Meerpohl[o,p], Gunn Vist[b], Elise Berliner[q], Susan Norris[r], Yngve Falck-Ytter[r], Holger J. Schünemann[a]

[a]*Departments of Clinical Epidemiology and Biostatistics and Medicine, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[b]*Norwegian Knowledge Centre for the Health Services, PO Box 7004, St. Olavs plass, 0130 Oslo, Norway*
[c]*Department of Medicine, Division of Gastroenterology, Hepatology, and Nutrition, University of Florida, Gainesville, FL, USA*
[d]*University of Oxford, Oxford, United Kingdom*
[e]*Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Barcelona 08041, Spain*
[f]*QUERI Program, Office of Research and Development, Department of Veterans Affairs, Washington, DC, USA*
[g]*Academy of Swiss Insurance Medicine (asim) University Hospital Basel Petergraben 4 CH-4031, Basel, Switzerland*
[h]*The Basel Institute of Clinical Epidemiology, University Hospital Basel Hebelstrasse 10, 4031 Basel, Switzerland*
[i]*Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN, USA*
[j]*Department of Medicine, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[k]*Harvard Medical School, UpToDate, Boston, USA*
[l]*Department of Urology, University of Florida, Gainesville, FL, USA*
[m]*Department of Medicine, State University of New York at Buffalo, Buffalo, NY, USA*
[n]*German Cochrane Center, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, 79104 Freiburg, Germany*
[o]*Division of Pediatric Hematology and Oncology, Department of Pediatric and Adolescent Medicine, University Medical Center Freiburg, 79106 Freiburg, Germany*
[p]*Technology Assessment Program, Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, USA*
[q]*Oregon Health and Science University, Department of Medical Informatics and Clinical Epidemiology, Portland, OR 97239-3098, USA*
[r]*Division of Gastroenterology, Case and VA Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA*

Accepted 15 January 2012; Published online 27 April 2012

Abstract

GRADE requires guideline developers to make an overall rating of confidence in estimates of effect (quality of evidence—high, moderate, low, or very low) for each important or critical outcome. GRADE suggests, for each outcome, the initial separate consideration of five domains of reasons for rating down the confidence in effect estimates, thereby allowing systematic review authors and guideline developers to arrive at an outcome-specific rating of confidence. Although this rating system represents discrete steps on an ordinal scale, it is helpful to view confidence in estimates as a continuum, and the final rating of confidence may differ from that suggested by separate consideration of each domain.

An overall rating of confidence in estimates of effect is only relevant in settings when recommendations are being made. In general, it is based on the critical outcome that provides the lowest confidence.

Keywords: GRADE; Quality of evidence; Confidence in estimates; Guideline methodology; Systematic review methodology; Values and preferences

1. Introduction

In prior studies in this series devoted to exploring GRADE's approach to rating confidence in estimates of effect (quality of evidence) and grading strength of recommendations (guidance for practice) we have dealt with issues of framing the question [1]; introduced GRADE's

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. Gordon H. Guyatt, CLARITY Research Group, Department of Clinical Epidemiology & Biostatistics, Room 2C12, 1200 Main Street West Hamilton, ON, L8N 3Z5, Canada. Tel.: +905-525-9140; fax: +905-523-8781.

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

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2012.01.006

What is new?

Key points

GRADE requires a rating of confidence in effect estimates (quality of evidence) for each outcome.

Rating of confidence of evidence requires a gestalt that simultaneously considers all eight domains (risk of bias, precision, consistency, and so forth)

Guideline developers using GRADE will subsequently make an overall rating of confidence in effect estimates across all outcomes based on those outcomes they consider critical to their recommendation.

Optimal application of GRADE requires making the reasons for key judgments transparent.

conceptual approach to rating the confidence in a body of evidence [2]; and presented five reasons for rating down the confidence in effect estimates (risk of bias [3], imprecision [4], inconsistency [5], indirectness [6], and publication bias [7]) and three reasons for rating up the confidence in effect estimates [8] (a large magnitude of effect, a dose-response gradient, and a situation in which plausible biases, if present, would serve to increase our confidence in the effect estimate), as well as dealing with issues specific to resource use. This 11th article in the series will focus on (1) summarizing the confidence in effect estimates across a single outcome for each important or critical outcome and (2) determining the confidence in effect estimates across all critical outcomes.

2. Summarizing the confidence in effect estimates for individual outcomes

GRADE's approach to rating down (or not) with respect to each of five criteria and to rating up (or not) with respect to three others is sometimes straightforward and enhances the transparency of the system. Most commonly, authors will be comfortable with the rating of confidence in estimate of effect that results from considering each criterion separately. Not infrequently, however, if ratings are applied in a blanket or rote fashion without considering context and the relation of one criterion to another, the confidence rating could be problematic. Specifically, ratings of individual domains could result in an overall rating of confidence in effect estimates on a particular outcome that does not correspond well to an integrated assessment or the gestalt of confidence in estimates of effect. In such instances, an adjustment in the final rating based on that gestalt is required.

Consider a systematic review of randomized trials of flavonoids for the treatment of hemorrhoids that produced a pooled estimate of a relative risk of persisting symptoms (lack of improvement) of 0.42 (95% confidence interval [CI] 0.28−0.61) [9]. Table 1 presents an evidence profile summarizing the evidence regarding two outcomes: persisting symptoms and adverse effects of the intervention. The profile presents the number of studies and patients, considerations related to the five possible reasons for rating down confidence in effect estimates (summarized in the table with expansions in the associated footnotes), and the best estimates and CIs around relative and absolute effects.

Consider now the possible reasons for rating down confidence in effect estimates. In most studies, the published articles left uncertainty whether allocation was concealed (though blinding in most suggests the likelihood of concealment), and all studies used unvalidated measures of symptoms. Given these limitations, one could reasonably argue either for or against rating down for risk of bias.

Fig. 1 presents a forest plot depicting the results of the review. The point estimates from individual studies are quite variable, and some of the CIs overlap little. The test for heterogeneity is highly significant and the I^2 large. All these observations suggest rating down for inconsistency among studies. On the other hand, all point estimates suggest benefit, and one might argue that it is inappropriate to rate down for inconsistency when the only uncertainty appears to be whether the magnitude of the treatment effect is moderate or very large. For instance, if undesirable consequences of an intervention are minimal, even a modest treatment effect may warrant a strong recommendation in favor of that treatment. If, in such a circumstance, the basis of doubt is whether the true effect is modest or large, rating down for inconsistency may well be inappropriate.

All available randomized trials were of small or moderate size (from 40 to 234 patients), and all were industry funded. This is a situation that raises the possibility of publication bias. In addition, one could interpret the funnel plot as suggesting the possibility of publication bias, with three small, very positive studies and no corresponding studies with small or negligible effects (Fig. 2). This line of reasoning would suggest rating down confidence in the estimate for publication bias. On the other hand, the number of studies is insufficient to meet rigorous criteria for creating a funnel plot [10] and one could argue that the case for publication bias is speculative in which case one would not rate down.

Thus, for three of the five domains in which one might rate down confidence in effect estimates (risk of bias, inconsistency, and publication bias) one could reasonably make the case for rating down or for not doing so. The situation is further complicated by the magnitude of effect: the relative risk of persisting symptoms (0.41) is slightly less than 0.5, raising the possibility of rating confidence up for the magnitude of effect. A generous reviewer, who in each case is inclined to view the results favorably, would interpret the body of evidence from these flavonoid studies as high quality (i.e., would not rate down the quality). A less generous reviewer, who decides to rate down the