Conclusiveness resolves the conflict between quality of evidence and imprecision in GRADE

Anttila, Sten; Persson, Johannes; Vareman, Niklas; Sahlin, Nils-Eric

Published in:
Journal of Clinical Epidemiology

DOI:
10.1016/j.jclinepi.2016.03.019

2016

Document Version:
Publisher's PDF, also known as Version of record

Journal of Clinical Epidemiology 75 (2016) 1–5

Journal of Clinical Epidemiology

VARIANCE & DISSENT - IS GRADE CONFUSING STATISTICAL TERMS AND SHOULD THEY USE THE TERM 'CONCLUSIVENESS' TO INFORM DECISION-MAKING?

Conclusiveness resolves the conflict between quality of evidence and imprecision in GRADE

Sten Anttila[a], Johannes Persson[b], Niklas Vareman[c,*], Nils-Eric Sahlin[c]

[a]*SBU: Swedish Council on Health Technology Assessment, Stockholm, Sweden*
[b]*Department of Philosophy, Lund University, Lund, Sweden*
[c]*Department of Medical Ethics, Lund University, Lund, Sweden*

Accepted 29 March 2016; Published online 5 April 2016

Abstract

Objectives: The objective of our article is to show how ''quality of evidence'' and ''imprecision,'' as they are defined in Grading of Recommendations Assessment, Development, and Evaluation (GRADE) articles, may lead to confusion. We focus only on the context of systematic reviews.

Study Design and Setting: We analyze, with the aid of standard probabilistic and statistical concepts, the concepts of quality of evidence and imprecision as used in the GRADE framework. This enables us to point out some weaknesses in the relation between ''quality of evidence'' and ''imprecision.''

Results: The GRADE framework contains terms familiar from classical statistics, but these terms are used in nonstandard ways. Notably, ''imprecision'' does not have the meaning in the GRADE framework that it has in statistics, and the well-known table of ''evidence levels'' wrongly suggests that ''quality of evidence'' and ''accuracy'' express the same concept—they do not.

Conclusion: We believe that ''conclusiveness'' rather than ''imprecision'' would be a suitable term to use when the question whether the CI excludes or includes certain critical margins is being addressed. Conclusiveness could also replace quality of evidence as the final step for a systematic reviewer.

Keywords: GRADE; Imprecision; Quality of evidence; Confidence in estimate; Accuracy; Conclusiveness

1. Introduction

Grading of Recommendations Assessment, Development, and Evaluation (GRADE) is a framework for rating quality of evidence and grading strength of recommendations. It is used as a tool for medicine, public health, and health policy. The ultimate aim of GRADE is to provide standardized clinical practice guidelines which address alternative management options (www.gradeworkinggroup.org/). Many major healthcare organizations around the world participate in the GRADE network, including WHO, Cochrane, AHRQ, and NICE.

GRADE is a valuable tool—there is no question about that. But does it, in its present form, work well enough to deliver on its promises? And does it provide adequate support where the rating and standardization of evidential ratings of intervention effects are concerned? These questions, we believe, are especially interesting given that the GRADE working group itself strongly recommends that there should be a single, unified specification of GRADE because [1].

> …modifications may confuse some users of evidence summaries … and because such changes compromise the goal of a single system with which clinicians, policy makers, and patients can become familiar. (p. 391)

In an effort to ensure that GRADE is not compromised, a number of guideline articles [2] have been published in the *Journal of Clinical Epidemiology* (JCE) (www.guidelinedevelopment.org/handbook/). Yet, local modifications do occur. Indeed, through personal communication with Professor Gordon Guyatt, we understand that the

Conflict of interest: There are no conflicts of interest for any author of the article.

* Corresponding author. Tel.: +46 46 222 09 06.
E-mail address: niklas.vareman@med.lu.se (N. Vareman).

http://dx.doi.org/10.1016/j.jclinepi.2016.03.019

GRADE group takes an active interest in conceptualizations of "quality of evidence."

The "purpose" of our article is to show how "quality of evidence" and "imprecision," as they are defined in GRADE articles, may lead to serious confusion. We focus only on the context of systematic reviews. We hope our contribution will be useful in further conceptualizing GRADE's definition of quality of evidence.

2. The GRADE process

In GRADE, quality of evidence is not determined via an algorithm designed to calculate specific evidential values. GRADE is intended to facilitate a systematic and transparent process. Crucial decisions are documented throughout this process, and this allows those considering the evidence to assess each step and decide whether to accept or reject the final rating. Unanimous agreement on the final rating is not guaranteed, however, and transparency seems to be just as important as consistency in quality rating [3].

The crucial decisions just mentioned can rate the level of evidence presented in a study down or up. Study design determines the initial rating in the process: lack of randomization in nonrandomized studies will lead to default downgrading by two levels unless no downgrading can be clearly and transparently justified by the fact that other measures to protect against bias and confounding that resemble randomization were observed. According to GRADE, this initial rating may be changed for any of eight specific reasons, five of which lower the rating and three of which raise it [3]. Rating down may follow from (1) study limitations, (2) inconsistency, (3) indirectness, (4) imprecision, or (5) publication bias. Rating up primarily concerns non-randomized studies. It may reflect (6) the magnitude of an effect, (7) the dose-response gradient, or (8) the fact that all plausible confounding would either reduce the demonstrated effect or increase the effect if no effect was observed. The outcome of this process, for the systematic reviewer, is a quality rating for every single outcome: high, moderate, low, or very low.

3. Quality of evidence

According to GRADE, quality of evidence in a systematic review is defined as follows [3]: "…the ratings of the quality of evidence reflect the extent of our confidence that the estimates of the effect are correct." It is obvious that quality of evidence is a property of the beholders, the systematic reviewers, and not of the research results as such. Furthermore, that quality (or level of confidence) is understood as a continuous variable [4] but is communicated on a four-part ordinal scale (Table 1).

The verbal expressions in Table 1 suggest that a "correct estimate" is understood as an estimate that "lies close" to the "true effect" (the parameter value). The quality of evidence, therefore, appears to be the reviewers' degree of confidence in the closeness of a parameter value to an estimated value.

Table 1. Elaboration of the four levels of evidence (modified table from [3], p. 404, our emphases)

Quality level	Definition
High	We are very confident that the "true effect lies close to that of the estimate of the effect"
Moderate	We are moderately confident in the effect estimate: The "true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different"
Low	Our confidence in the effect estimate is limited: The "true effect may be substantially different from the estimate of the effect"
Very low	We have very little confidence in the effect estimate: The "true effect is likely to be substantially different from the estimate of the effect"

4. Imprecision

Imprecision is one of five reasons for rating down evidential quality [5]. For systematic reviewers, the concept of imprecision includes several components: absolute sample size, a retrospective statistical power calculation (p. 5) called optimal information size (OIS), confidence intervals (CI), and critical margins regarding "no effect," "important benefit," and "important harm." This is expressed in Table 2 as a rule for rating down quality of evidence.

GRADE imprecision is thus a combination of more than one aspect of statistical power, confidence intervals, and specified limits (in other words, critical margins).

5. The problem

The problem is that "imprecision" includes a component that obviously does not affect the closeness of the parameter value to the estimate—namely, assessment of whether the CI excludes certain critical margins. This problem is demonstrated in Fig. 1.

Assume that the OIS criterion is met for both results A and B. A excludes "no effect," whereas B includes it. Therefore, the systematic reviewer should rate down for imprecision in case B, but not in case A.

Table 2. Rule for rating down—systematic reviews (modified from [5] p. 4, our emphases)

If the optimal information size [OIS] criterion is not met, rate down for imprecision, unless the sample size is very large (n $\geq$ 2,000 [The limit "200 or perhaps 4,000 patients" is stated in the GRADE article [5], but we believe that 2,000 patients is the intended number.] to 4,000 patients [It is not clear whether the criterion for exclusion and nonexclusion of critical margins (including "no effect") applies when N $\geq$ 2,000 (or 4,000) is met.])…
If the "OIS criterion" is met and the 95% CI excludes no effect… precision adequate
If OIS is met, and CI overlaps no effect… rate down if CI fails to exclude "important benefit or important harm"

Abbreviation: CI, confidence interval.

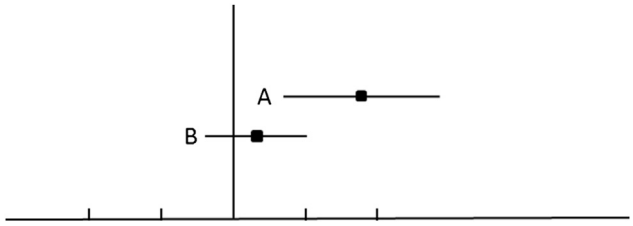

Fig. 1. Inclusion of "no effect" (hypothetical results). (Point estimate is symbolized by a black square; CI is symbolized by a vertical line.)

After the GRADE definition of quality of evidence, the systematic reviewer should be very confident that the parameter value in case A lies close to the estimate. In case B, the reviewer should be (possibly) moderately convinced that the parameter value lies close to the estimate—in other words, he or she should be less confident in case B than in case A. In both cases, we assume that there are no other reasons for rating down the quality of evidence. Now, observe that the CI is longer in case A than it is in B. It is obviously confusing that the reviewer should have more trust in the closeness in case A than he or she does in B.

The introduction of "important benefit" and "important harm" gives rise to the same kind of problem. In Fig. 2, there are four hypothetical results (C to F). We symbolize "margin of important harm" by Δ^- and "margin of important benefit" by Δ^+. Within region R, the effects are so small that they are of no clinical relevance.

Assume that the OIS criterion is met in all of C to F. The precision of C is adequate already at step 2 in Table 2, since "no effect" is excluded. Results D, E, and F, on the other, hand are candidates for rating down. Since D and E exclude Δ^-, precision here is adequate according to GRADE. F, however, is rated down on the basis of imprecision (Table 2). The CI for C and for D is longer than it is for F. Does this mean that the systematic reviewer should be more convinced that the estimate in C and D lies close to the parameter value than her or she is in case F? Confusion may arise as to what measure to use.

6. Conceptual sources of confusion

The GRADE framework contains terms familiar from classical statistics, but these terms are used in nonstandard ways. Notably, "imprecision" does not have the meaning in the GRADE framework that it has in statistics. Adding to the confusion, the concept of imprecision deployed in the GRADE framework is not well defined. Another problem is that the well-known table of "evidence levels" wrongly suggests that "quality of evidence" and "accuracy" express the same concept—they do not.

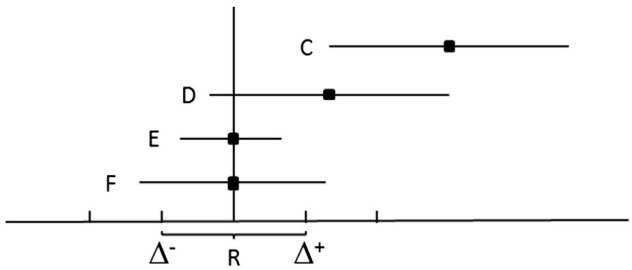

Fig. 2. Inclusion of important "benefit" and "harm" (hypothetical results).

In statistics [6], "accuracy" is a familiar concept expressing closeness between a parameter value and an estimate; it encompasses both bias and sample precision (p. 267–9). This can be illustrated by using the idea of an arithmetic mean. Thus, bias, in this particular case, is expressed as the difference between the expected sample mean and population mean (Equation 1).

$$bias = E(\bar{x}) - \mu_x \tag{1}$$

Sample precision is then the variance of the observations about the sample mean (Equation 2).

$$sample\ precision = s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2}$$

The combination of sample precision and bias can, accordingly, be calculated as in Equation 3 (provided that there is available information indicating bias).

$$accuracy = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - (\bar{x} - bias))^2 \tag{3}$$

Finally, accuracy can be simplified as in Equation 4 by substituting μ_x for $(\bar{x} - bias)$.

$$accuracy = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \mu_x)^2 \tag{4}$$

In this example, bias, precision, and accuracy are all explicitly expressed, and the relationships between the concepts are clear.

The notion of quality of evidence formulated in GRADE guideline articles is reminiscent of accuracy. The closeness of a parameter value to an estimate is clearly indicated in Table 1 above, but the quality of evidence is not explicitly defined, so we cannot be sure what it is. GRADE imprecision is more difficult to understand from a statistical point of view, since it is based on 3–4 components, as we described previously. The content of the concepts is verbally indicated by the decision rule stated in Table 2. The relationship of "imprecision" to "quality of evidence" is unfortunately not explicitly explained.

In practice, we are often faced with a CI calculated on the assumption that there is no bias present, although we suspect that the sample mean $\bar{x}$ and the sample variation s_x^2 are biased to some degree. A biased $\bar{x}$ implies that the point estimate is too high or too low in comparison to μ_x, and also that the limits of the CI are either too high or too low. A biased s_x^2 entails that the CI is too long or too short [CI is $\bar{x} - t_{\alpha/2} \cdot s_{\bar{x}} \leq \mu_x \leq \bar{x} + t_{\alpha/2} \cdot s_{\bar{x}}$. Length depends

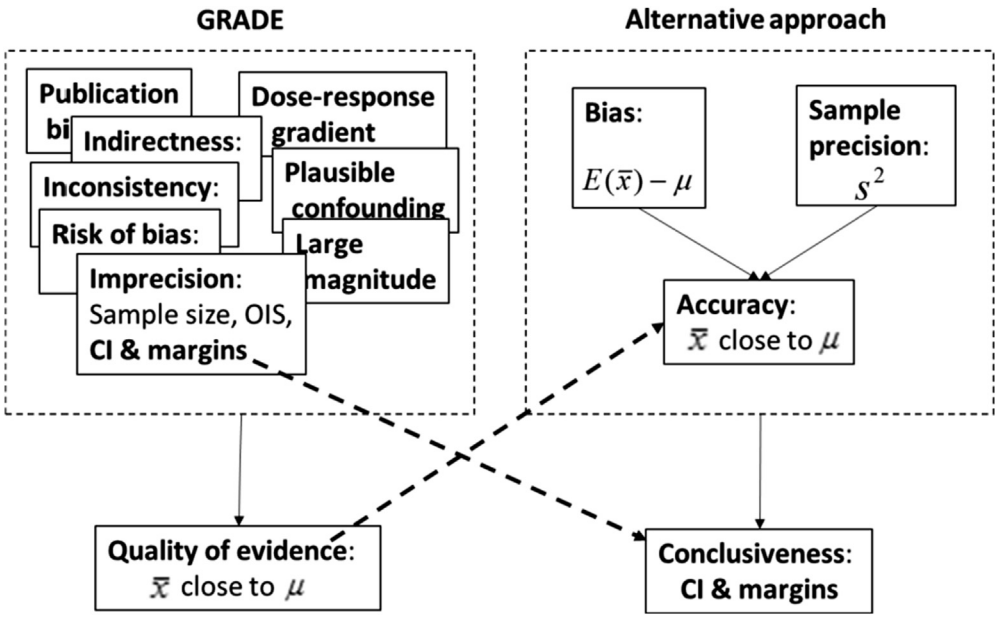

Fig. 3. Two alternative conceptual approaches. CI, confidence interval; OIS, optimal information size.

on $s_{\bar{x}}$ and α(choice based). $s_{\bar{x}}$ determined by n and to s_x due to $s_{\bar{x}}=\sqrt{s_x^2/n}$. Calculation assumes $E(s_x^2)=\sigma_x^2$. Possible bias $E(s_x^2 \neq \sigma_x^2)$.]. This means that there are uncertainties not captured by the CI.

We believe the GRADE framework has great potential when we try to understand and assess biases in $\bar{x}$ and s_x^2. The reasons for rating down or up address such biases more or less clearly, but explicit definitions and specifications are needed.

If we have reasonable knowledge of the size and direction of bias, the CI can be adjusted accordingly. If we do not, the identified biases can be treated as uncertainty factors that may either be included in a model [7] or approached using the GRADE rates: rate 4 for minimal uncertainty and rate 1 for maximal uncertainty.

7. A procedural source of confusion

We think that critical margins are included in imprecision too early in the evaluation process presented in contemporary GRADE. It would be preferable for the comparison of a CI with critical margins to be preceded by the assessment of accuracy.

In Fig. 3, the GRADE approach is compared with an alternative. In GRADE, confidence in the closeness of a parameter value to an estimate (quality of evidence) is preceded by a consideration of the eight possible reasons for rating down or up, including imprecision (with the inclusion or exclusion of critical margins). In this alternative approach, closeness (accuracy) is based on bias and precision. Once accuracy has been determined, the (possibly) adjusted CIs are compared with the critical margins. This is a standard procedure for dealing with uncertainties (See for example, ISO 17025 [www.iso.org/iso/home.html] and Guide to the Expression of Uncertainty in Measurement [GUM; www.bipm.org/en/publications/guides/gum.html].) and the compliance with quality margins specified by decision makers and stake holders; a common term for such margins is ‘‘tolerance frames’’ or levels [8]. First, there is an evaluation—and statement—of uncertainty; there follows a comparison of adjusted CIs with tolerance levels.

8. Suggestions

We think ‘‘conclusiveness’’ rather than ‘‘imprecision’’ would be a suitable term to use when the question whether the CI excludes or includes certain critical margins is being addressed. Conclusiveness could also replace quality of evidence as the final step for a systematic reviewer.

This suggestion is in line with the conceptual framework of noninferiority and equivalence trials [9]. In noninferiority trials, the critical noninferiority margin is introduced, not as an aspect of imprecision, but as a support for decision making. When noninferiority margin is included, the result is called inconclusive. When it is not, the result is conclusive and either noninferior or inferior.

Finally, ‘‘accuracy’’ is a better term than ‘‘quality of evidence’’ when closeness is being addressed.

9. Concluding remarks

Our analysis suggests that in the GRADE guideline articles the key notions of ‘‘quality of evidence’’ and ‘‘imprecision’’ are, independently, but especially when taken

together, a source of serious confusion, and that this may impede the practical process of evidence formation. Explicit GRADE guidance is therefore required here.

Several other issues in the GRADE guidance need to be discussed, but not here. Possibly the most important issue concerns the difference between evaluating evidence in a systematic review and giving guidelines [10], and specifically the question how an evidential value is to be transformed into a value used for regulatory purposes.

A separate issue arises because second-order uncertainty seems to be introduced in the GRADE approach [11,12] as a result of ‘‘quality of evidence’’ being interpreted in GRADE as referring to the assessor’s epistemic confidence in the CI. It is important to distinguish between different types of uncertainty occurring at different stages in the GRADE process. A third issue focuses on the translation of quality of evidence from a continuous to an ordinal level variable [13].

Acknowledgments

This work is supported by a grant from the Swedish Foundation for Humanities and Social Sciences, grant number M14-0138:1. Personal communication between the authors and Professor Gordon Guyatt, of McMaster University, Hamilton, Ontario (CA), was of great help in the preparation of the article. The authors are grateful for valuable comments from Susanna Axelsson, Jenny Odeberg, Natalie Peira, and Sigurd Vitols SBU, and from the editors of the Journal of Clinical Epidemiology. Thanks are due also to colleagues in the VBE consortium who commented on earlier drafts of this article. The members of the VBE consortium (apart from the present authors) are: Wändi Bruine de Bruin, Leeds University Business School (UK), Johan Brännmark, Malmö University (Sweden), Alex Davis, Carnegie Mellon University (US), Baruch Fischhoff, Carnegie Mellon University (US), Charlotta Levay, Lund University (Sweden), Barbara McNeil, Harvard Medical School (US), Lena Wahlberg, Lund University (Sweden), Annika Wallin, Lund University (Sweden).

References

[1] 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;64:12.

[2] Guyatt G, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011;64:3.

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

[4] Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. J Clin Epidemiol 2013;66:7.

[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;64:11.

[6] Ayyub BM, McCuen RH. Probability, Statistiscs, and Reliability for Engineers and Scientists. Boca Raton, London, New York: CRC Press, Taylor Francis Group, A Chapman & Hall Book; 2011.

[7] Taylor JR. An Introduction to Error Analysis. The Study of Uncertainties in Physical Measurement. Sausalito, CA, USA: University Science Books; 1997.

[8] Auty FJ, Bevan K, Hanson A, Machin G, Scott J, Brown C, Haritos G, Martinez-Botas RF. Beginner’s guide to Measurement in Mechanical Engineering 2014:Good Practice Guide No.131. National Physical Laboratory (NPL), Teddington & London.

[9] Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA 2012;308:11.

[10] Vareman N, Persson J. Why separate risk assessors and risk managers? Further external values affecting the risk assessor qua risk assessor. J Risk Res 2010;13(5):14.

[11] Gärdenfors P, Sahlin N-E. Unreliable probabilities, risk taking, and decision making. Synthese 1982;53(3):26.

[12] Sundgren D, Karlsson A. Uncertainty levels of second-order probability. Polibits 2013;(48):12.

[13] Kampen J, Swyngedouw M. The ordinal controversy revisited. Qual Quantity 2000;34:16.