OPEN ACCESS

When applying GRADE, how do we decide the target of certainty of evidence rating?

Linan Zeng ,[1,2] Romina Brignardello-Petersen,[1] Gordon Guyatt[1]

[1]Health Research Methods, Evidence & Impact, McMaster University, Hamilton, Ontario, Canada
[2]Pharmacy Department/ Evidence-based Pharmacy Center, West China Second University Hospital, Sichuan University, Chengdu, China

Correspondence to
Dr Linan Zeng, McMaster University, Hamilton, Canada; zengl15@mcmaster.ca

Received 8 February 2021
Accepted 15 May 2021

ABSTRACT

The Grades of Recommendation, Assessment, Development and Evaluation' (GRADE) offers a widely adopted, transparent and structured process for developing and presenting summaries of evidence, including the certainty of evidence, for systematic reviews and recommendations in healthcare. GRADE defined certainty of evidence as 'the extent of our confidence that the estimates of the effect are correct (in the context of systematic review), or are adequate to support a particular decision or recommendation (in the context of guideline)'. Realising the incoherence in the conceptualisation, the GRADE working group re-clarified the certainty of evidence as 'the certainty that a true effect lies on one side of a specified threshold, or within a chosen range'. Following the new concept, in the context of both systematic reviews and health technology assessments, it is desirable for GRADE users to specify the thresholds and clarify of which effect they are certain. To help GRADE users apply GRADE in accordance with the new conceptualisation, GRADE defines three levels of contextualisation: minimally, partially and fully contextualised approaches, and provides possible thresholds for each level of contextualisation. In this article, we will use a hypothetic systematic review to illustrate the application of the minimally and partially contextualised approaches, and discuss the application of a fully contextualised approach in deciding how we are rating our certainty (i.e.target of the rating of certainty of evidence).

The Grades of Recommendation, Assessment, Development and Evaluation (GRADE) approach, for categorising certainty of evidence (also referred to quality of evidence, confidence in the evidence) and strength of recommendations, is increasingly adopted by systematic reviewers, health technology evaluators and guideline panels worldwide. The GRADE approach has advantages over previous rating approaches in several aspects, including a clear separation and explicit definition of certainty of evidence and strength of recommendations.[1]

GRADE initially provided formal definition of certainty of evidence as 'the extent of our confidence that the estimates of the effect are correct' in the context of systematic reviews and health technology assessment, which suggests that we are rating our certainty in point estimates of effect, or 'the extent of our confidence that the estimates of an effect are adequate to support a particular decision or recommendation' in the context of making recommendations, which suggests that we are rating certainty that the true effect is on one side or another of a decision threshold.[2] Realising the incoherence in its previous conceptualisation, in 2017 the GRADE working group provided an alternative definition of certainty of evidence.[3]

In this paper, we will review the reasoning that led to modifying the former concept and adopting the current one, and illustrate how the new conceptualisation applies to what GRADE has called degrees of contextualisation.

THE FORMER AND CURRENT CONCEPTUALISATIONS

In the GRADE approach, randomised control trials (RCTs)—the focus of this discussion—begin as high certainty and may be rated down to moderate, low or even very low certainty as a result of concerns of risk of bias, inconsistency, imprecision, indirectness and publication bias. In our current discussion, we will use a hypothetical systematic review in mental health and focus on the GRADE domain of imprecision. The hypothetical systematic review of antidepressants versus placebo for the treatment of adults with major depressive disorder based on RCTs yielded a pooled estimate of absolute increase in response rate of 6.3%, with 95% CI of 3.6%–9.0% (Response rate was measured by the total number of patients who had a reduction of ≥50% of the total score on a standardised scale for depression) (figure 1).

In GRADE approach, the judgement of imprecision refers to whether the CI of around the estimate of effect is sufficiently narrow that one need not rate down for imprecision, or sufficiently wide that one should. It turns out to be impossible to make this decision without referring to some sort of threshold for 'too wide'.[4] Consider the decision of whether to rate down for imprecision in figure 1. What threshold might we choose? If we choose no effect as a threshold and rate our certainty in relation to the threshold, we would rate our certainty that the true effect of antidepressants is larger than no effect (a non-null effect presents). Since the 95% CI does not overlap with this threshold (no effect), we would not rate down for imprecision. If there were no other concerns in any GRADE domains, the certainty of this evidence would be high.

Alternatively, we might specify that a response rate increase of less than 5 in 100 represents an effect too small and unlikely to be important to patients (ie, 5 in 100 represents a minimally important difference, ie, the MID) and rate our certainty that antidepressants result in a greater than trivial effect. Now, the CI overlaps this alternative threshold and we would rate down our certainty for imprecision.

To cite: Zeng L, Brignardello-Petersen R, Guyatt G. *Evid Based Ment Health* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/ebmental-2020-300170

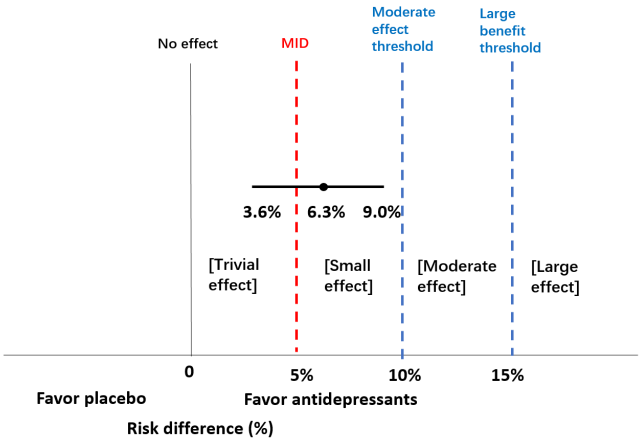

Figure 1 A hypothetic systematic review of antidepressants versus placebo for the treatment of adults with major depressive disorder based on RCTs. The pooled estimate: 6.3 more in 100 patients response in antidepressants group, 95% CI : from 3.6 more to 9 more in 100 patients. Minimally important difference (MID): 5 more in 100 patients respond. Moderate effect threshold: 10 more in 100 patients respond. Large effect threshold: 15 more in 100 patients respond. RCTs, randomised control trials.

In this case, if there were no other concerns in any GRADE domains, the certainty of this evidence would be moderate.

This logic, which applies to all ratings of imprecision, led GRADE to the new characterisation of certainty ratings. GRADE now considers the certainty of evidence as our confidence that a true effect lies on one side of a specified threshold, or within a chosen range either in the context of systematic review or in the context of making recommendations.[3]

The hypothetical example illustrates that when using GRADE for certainty ratings, it is highly desirable to specify the thresholds or ranges and clarify what it is in which we are rating our certainty (ie, the target of certainty rating). To help GRADE users choose the possible thresholds or ranges, GRADE defines three levels of contextualisation, minimally, partially and fully contextualised, and provides possible thresholds for each level of contextualisation.[3]

MINIMALLY CONTEXTUALISED AND PARTIALLY CONTEXTUALISED APPROACHES

The minimally and the partially contextualised approaches are relevant primarily for systematic reviews and health technology assessments. In a minimally contextualised approach, possible thresholds are no effect, or the MID. The MID represents the smallest change in the particular outcome that patients perceive as important.[5] In the hypothetical systematic review, if we applied a minimally contextualised approach, we could—as we have done above—either rate our certainty in relation to no effect, or set a threshold of MID that represents a small but important increase in response rate (in this case an 5 in 100 or 5% increase in response) and rate our certainty in relation to the MID. As we have illustrated, choosing the threshold of no difference would lead us to rating our certainty that the true effect of antidepressants is larger than no effect and not rating down for imprecision, while choosing the MID would lead us to rating our certainty that the true effect of antidepressant is larger than a minimally important effect and rate down for imprecision (figure 1). Note, it is possible to set the thresholds in the minimally contextualised approach without referring to outcomes other than response.

Using a partially contextualised approach, we still set thresholds for a single outcome without considering other outcomes. Here, we rate our certainty that the true effect lies within a specified range that represents a particular magnitude of effect. GRADE recommends four ranges (trivial, small, moderate or large effect) divided by three thresholds (MID, moderate effect threshold, large effect threshold) (figure 1). Note, when rating certainty in relation to threshold(s) other than no effect, we should decide and present all other thresholds and effect estimates in absolute terms.

Returning to the hypothetical example, using a partially contextualised approach, if we set the moderate effect threshold at 10%, and large effect threshold at 15% increase in response, we could rate our certainty that the true effect lies within the range of small effect (ie, a small but important effect is present) and would rate down for imprecision due to the overlap of the 95% CI with the MID (figure 1). We could also rate our certainty in relation to the moderate effect threshold (ie, the true effect is smaller than a moderate effect) and we would not rate down for imprecision. If we chose to rate our certainty in relation to the large effect threshold (ie, the true effect is smaller than a large effect), because the 95% CI is entirely below the large effect threshold, we would not rate down for imprecision.

When more than one possible threshold is available, the choice of what to rate the certainty in depends on what is most useful to the target audience (eg, clinicians managing patients with major depressive disorder). It might be desirable to choose more than one option of what it is in which to rate our certainty for a particular outcome when multiple ratings are useful to the target audience. For instance, a clinician might find it useful to inform a patient that a treatment effect is most likely to represent a small but important difference (the point estimate), but could be even smaller (below the MID), but is certainly less than a large effect (below the large effect threshold). Such statements, however, require that patients view the MID, and moderate and large thresholds, as depicted in figure 1.

FULLY CONTEXTUALISED APPROACH

Fully contextualised approaches, typically used in guidelines, require simultaneously considering all critical outcomes, associated values and preferences judgements, and setting a threshold for the net effect (defined as benefits minus harms). The process is challenging, but the need is clear.[3]

When one can specify a single key benefit outcome, the approach is straightforward (at least for that outcome). Considering all harms, one can decide on a threshold of magnitude of benefit that would warrant administration of the intervention. If the entire CI is below the threshold, the intervention is not warranted. If the entire CI is above the threshold, a guideline panel can recommend the intervention. If the CI crosses the threshold, one would rate down for imprecision.

A recent BMJ Rapid Recommendation addressing colorectal cancer screening has shown the feasibility of this approach being applied to guide a formal recommendation.[6 7] The guideline panel first identified reduction in colorectal cancer incidence and mortality as the key benefit outcome of cancer screening. When presented with the evidence on harms and burdens associated with cancer screening, the guideline panel gave their views of the expected benefits on colorectal cancer incidence and mortality that people would require to undergo screening, given the harms and burdens during a survey. Then, the panel discussed the results

of the surveys and agreed on thresholds for benefits at which the majority of people would choose screening.[7] This process greatly facilitated an explicit specification of the threshold for the magnitude of key benefit required to justify screening given the harms and burdens. The approach is again straightforward if there is one key harm outcome. Here, the logic is identical, except that one considers all the benefits and then specifies a threshold magnitude of harm that would warrant not using the intervention.

GRADE also offers other approaches in which multiple benefit outcomes and multiple harms and burdens are considered simultaneously. One approach starts with selecting one outcome as a reference outcome and defining a relative importance for each other outcome and then deciding the direction of recommendation by calculating the weighted net effect.[3] The 2017 conceptual paper presented an example of this approach using the decision regarding whether to use shorter or longer duration of dual antiplatelet therapy in patients who have undergone placement of drug eluting stents in their coronary arteries.[3]

Another approach to addressing certainty of evidence in the fully contextualised is to generate a net effect estimate, defined as the certainty that the balance between desirable and undesirable health effects is favourable (ie, net benefit) or unfavourable (ie, net harm), and its CIs by applying decision analysis.[8] Algorithm-supported calculators facilitate combining the importance-adjusted effect estimates and classifying the precision.[8] This alternative approach, however, involves conceptual and computational challenges that, to our knowledge, guideline panels have not yet taken on.

RATING THE CERTAINTY OF THE EVIDENCE DOES NOT ALWAYS REQUIRE SPECIFYING EXACT THRESHOLD(S)

In general, when people are presented with a particular magnitude of effect, they find it relatively easy to say whether that effect is or is not important. They generally find this much easier than specifying the threshold that divides an important effect from an unimportant effect. It may sometimes, therefore, be useful to ask people (a guideline panel or systematic review team, or yourself) to specify whether the boundaries of a CI are on the same side of a threshold—an implicit way of gaining insight into the threshold—rather than asking them to choose a single threshold. To put it another way, it is reasonable to say we are rating our certainty that an important effect exists without, at the outset, specifying the single threshold that represents the boundary between an important and an unimportant effect.

Undoubtedly, there are dangers to this alternative—preconceptions regarding the usefulness of an intervention might influence judgements about whether effects are important, and therefore the implicit placement of thresholds. The a priori specification of thresholds is therefore ideal, and our innovative approach to aid guideline panellists to make that decision may prove widely useful.[7] There is, nevertheless, a tradeoff with the complexity of the judgements with which one burdens guideline panellists: having considered the risks of preconceptions influencing implicit threshold judgments, guideline steering groups may still reasonably consider the implicit judgement approach.

It is clear both that there are challenges to applying certainty of evidence in the fully contextualised setting, and that explicit values and preferences judgements are required to do so. Although many guideline panels have successfully taken on the challenge of making values and preferences explicit and then using fully contextualised approaches to judging certainty of evidence, they still represent only a small minority.[9]

CONCLUSION

GRADE defines certainty of evidence as the certainty that a true effect lies on one side of a specified threshold, or within a chosen range. When using GRADE, it is desirable to specify the thresholds or ranges and clarify the target of certainty of evidence rating. GRADE provides three levels of contextualisation, minimally, partially and fully contextualised approach, and possible thresholds under each level of contextualisation to facilitate certainty ratings.

Contributors GG, RB-P and LZ conceived the study. LZ drafted the manuscript. GG and RB-P helped refine the manuscript. All authors have read and approved the final version.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests LZ, RB-P and GG are GRADE working group members. GG is the co-founder of GRADE working group. Although consistent with GRADE guidance, this article does not constitute official guidance from the GRADE working group.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD

Linan Zeng http://orcid.org/0000-0001-9892-2000

REFERENCES

1 Guyatt GH, Oxman AD, Vist GE, *et al*. Grade: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924–6.
2 Balshem H, Helfand M, Schünemann HJ, *et al*. Grade guidelines: 3. rating the quality of evidence. *J Clin Epidemiol* 2011;64:401–6.
3 Hultcrantz M, Rind D, Akl EA, *et al*. The grade Working group clarifies the construct of certainty of evidence. *J Clin Epidemiol* 2017;87:4–13.
4 Guyatt GH, Oxman AD, Kunz R, *et al*. GRADE guidelines 6. Rating the quality of evidence--imprecision. *J Clin Epidemiol* 2011;64:1283–93.
5 Guyatt GH, Osoba D, AW W. Clinical significance consensus meeting group. methods to explain the clinical significance of health status measures. *Mayo Clinical Proceedings* 2002;77:371–83.
6 Helsingen LM, Vandvik PO, Jodal HC, *et al*. Colorectal cancer screening with faecal immunochemical testing, sigmoidoscopy or colonoscopy: a clinical practice guideline. *BMJ* 2019;367:l5515.
7 Helsingen LM, Zeng L, Siemieniuk RA, *et al*. Establishing thresholds for important benefits considering the harms of screening interventions. *BMJ Open* 2020;10:e037854.
8 Alper BS, Oettgen P, Kunnamo I, *et al*. Defining certainty of net benefit: a grade concept paper. *BMJ Open* 2019;9:e027445.
9 Zeng L, Helsingen LM, Kenji Nampo F, *et al*. How do cancer screening guidelines trade off benefits versus harms and burdens of screening? A systematic survey. *BMJ Open* 2020;10:e038322. doi:10.1136/bmjopen-2020-038322