Journal of Clinical Epidemiology 150 (2022) 216–224

Journal of Clinical Epidemiology

GRADE GUIDANCE SERIES

GRADE Guidance 34: update on rating imprecision using a minimally contextualized approach

Linan Zeng[a,b,*], Romina Brignardello-Petersen[b], Monica Hultcrantz[c], Reem A. Mustafa[d], Mohammad H. Murad[e], Alfonso Iorio[b,f], Gregory Traversy[g], Elie A. Akl[h], Martin Mayer[i,j,k], Holger J. Schünemann[b,f], Gordon H. Guyatt[b,f]

[a]*Pharmacy Department/Evidence-based Pharmacy Centre, West China Second University Hospital, Sichuan University and Key Laboratory of Birth Defects and Related Disease of Women and Children, Ministry of Education, No. 20, Section 3, South Renmin Road, Chengdu 610041, China*
[b]*Department of Health Research Methods, Evidence and Impact, McMaster University, 1280 Main Street West, Hamilton, L8S 4L8 Ontario, Canada*
[c]*Swedish Agency for Health Technology Assessment and Assessment of Social Services (SBU), S:t Eriksgatan 117, Stockholm 102 33, Sweden*
[d]*Division of Nephrology and Hypertension, Department of Internal Medicine, University of Kansas Medical Centre, 3901 Rainbow Blvd, MS3002, Kansas City, KS 61160, USA*
[e]*Evidence-based Practice Center, Mayo Clinic, 200 1st ST. SW, Rochester, MN 55905, USA*
[f]*Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, L8S 4L8 Ontario, Canada*
[g]*Public Health Agency of Canada, 785 Carling Avenue, Ottawa, K1A 0K9 Ontario, Canada*
[h]*Department of Internal Medicine, American University of Beirut, P.O.Box 11-0236, Beirut, Lebanon*
[i]*EBSCO Clinical Decisions, EBSCO, 10 Estes StIpswich, MA 01938, USA*
[j]*Triad Hospitalists, Cone Health, 1200 North Elm St, Greensboro, NC 27401, USA*
[k]*Open Door Clinic, Cone Health, 319 N Graham Hopedale Rd, Burlington, NC 27217, USA*

Accepted 25 July 2022; Published online 4 August 2022

Abstract

Objectives: The aim of this study is to provide updated guidance on when The Grading of Recommendations Assessment, Development and Evaluation (GRADE) users should consider rating down more than one level for imprecision using a minimally contextualized approach.

Study Design and Setting: Based on the first GRADE guidance addressing imprecision rating in 2011, a project group within the GRADE Working Group conducted iterative discussions and presentations at GRADE Working Group meetings to produce this guidance.

Results: GRADE suggests aligning imprecision criterion for systematic reviews and guidelines using the approach that relies on thresholds and confidence intervals (CI) of absolute effects as a primary criterion for imprecision rating (i.e., CI approach). Based on the CI approach, when a CI appreciably crosses the threshold(s) of interest, one should consider rating down two or three levels. When the CI does not cross the threshold(s) and the relative effect is large, one should implement the optimal information size (OIS) approach. If the sample size of the meta-analysis is far less than the OIS, one should consider rating down more than one level for imprecision.

Conclusion: GRADE provides updated guidance for imprecision rating in a minimally contextualized approach, with a focus on the circumstances in which one should seriously consider rating down two or three levels for imprecision.

Keywords: GRADE; Imprecision; Minimally contextualized approach; Systematic review; Guideline

Funding: L.Z. is funded by the Science and Technology Plan Project of Sichuan Province (2020YFS0035).

Conflict of interest: All authors of this paper are GRADE Working Group members. Gordon H. Guyatt and Holger J. Schünemann are the co-founders and co-chairs of GRADE Working Group.

Author Contributions: G.H.G., R.B.P., L.Z., M.H., and H.J.S. conceptualized this work and collected examples. L.Z., R.B.P., and G.H.G. drafted the manuscript. R.A.M., M.H.M., A.I., G.T., E.A.A., and M.M. provided input that resulted in important modifications. All authors approved the final manuscript.

* Corresponding author. Sichuan University and Key Laboratory of Birth Defects and Related Disease of Women and Children, West China Second University Hospital, No. 20, Section 3, South Renmin Road, Chengdu 610041, China. Tel.: +86-028-88570235; fax: +86-028-85503785.

E-mail address: zengl15@mcmaster.ca (L. Zeng).

https://doi.org/10.1016/j.jclinepi.2022.07.014

What is new?

Key findings

- In systematic reviews and guidelines The Grading of Recommendations Assessment, Development and Evaluation (GRADE) now suggests aligning approach that relies on thresholds and CIs of the absolute effect (i.e., CI approach) as a primary criterion for imprecision rating.
- Using the CI approach, when the CI is wide and considerably cross the threshold(s) of interests (i.e., one or both boundaries of CIs suggest inferences appreciably different from point estimate), one should consider rating down two levels for imprecision, and when the CI is very wide that the two boundaries of CI suggest very different inferences, one should consider rating down three levels for imprecision.
- Using the OIS approach, for dichotomous outcomes, one should consider rating down two levels for imprecision, when the ratio of the upper to the lower boundary of the CI is more than 2.5 for odds ratio or three for risk ratio; for continuous outcomes, when the sample size is smaller than 30–50% of the OIS.
- When the baseline risk is very low, GRADE suggests being more restrained in rating down for imprecision.

What this adds to what is known?

- Building on prior GRADE guidance on imprecision rating, this article provides specific guidance on circumstances when, using a minimally contextualized approach, one should consider rating down more than one level for imprecision based on the CI approach and the OIS approach.

What is the implication, what should change now?

- The article alerts GRADE users to the merits of using the same CI approach to imprecision in both systematic reviews and guidelines.
- GRADE users should seriously consider rating down imprecision by more than one level when the CI appreciably crosses the threshold(s) of interest; or when the CI does not cross the threshold(s) and the relative effect is large, the sample size of meta-analysis is far less than the OIS.

1. Introduction

The first specific The Grading of Recommendations Assessment, Development and Evaluation (GRADE) guidance addressing rating down certainty of evidence due to imprecision appeared in 2011 [1]. According to that guidance, GRADE's primary criterion for judging imprecision in guidelines focuses on the confidence interval (CI) around the absolute difference between intervention and control and relates the CI to a decision threshold. If the CI crosses the threshold, one rates down; if it does not cross, one does not (hereafter ''CI approach''). When relative effect is large and both sample size and number of events are modest, results may be fragile even if the CI appears satisfactorily narrow. In such instances, rating down for imprecision on the basis of the optimal information size (OIS) (also called review information size) may be appropriate (hereafter ''OIS approach'') [1]. For systematic reviews, the focus was on whether the sample size of meta-analysis meets the OIS (i.e., OIS approach) rather than judgments associated with thresholds (i.e., CI approach) [1]. If in guidelines the CI is very wide or in systematic reviews the OIS is far less than met that guidance suggests to consider rate down by two levels [1].

After almost a decade of additional experience, GRADE methodologists have had some important additional insights:

1. Systematic reviews are most useful for target audiences when authors make judgements in relation to possible thresholds. Thus, authors of systematic reviews are much more likely to use the approach that relies on thresholds and CIs around the absolute effect (i.e., CI approach) than OIS to judge imprecision (i.e., OIS approach).
2. When the CI appreciably crosses the threshold(s) of interest, authors of systematic reviews and guidelines should be more inclined to consider rating down more than one level (i.e., two or three levels) for imprecision.

Before making imprecision judgements, authors should be clear about the target of certainty rating (Box 1) [2]. In this paper, we will illustrate how, after determining the target of certainty rating, authors should assess imprecision in a minimally contextualized setting.

2. GRADE now suggests aligning imprecision criteria for systematic reviews and guidelines using confidence interval approach as a primary criterion for imprecision rating

The argument against the approach of using thresholds to make judgments about imprecision (i.e., CI approach) in systematic reviews is that setting thresholds involves value judgments that may be unfeasible for systematic reviewers, and that is central to decisions for guideline developers [1]. Our subsequent experience is that optimal systematic reviews inevitably involve some value judgments, and the notion of thresholds and imprecision ratings on that basis is intuitive and as a result useful to clinician audiences. When setting thresholds, systematic review

Box 1 Possible threshold(s) of interest and target of certainty rating in minimally, partially, or fully contextualized approach

Using a minimally contextualized approach (typically in systematic reviews), authors consider only one outcome at a time. Authors rate their certainty in relation to the null—rating their certainty that an effect is truly present—or in relation to a minimally important difference (MID)—rating their certainty that an important effect is truly present.

Using a partially contextualized approach, authors rate their certainty that the true effect falls in a range representing a trivial, small, moderate, or large effects for one outcome at a time.

In a fully contextualized approach (typically used in guidelines), authors simultaneously consider multiple outcomes (i.e., trading off desirable vs. undesirable health effects of an intervention) and set a decision threshold above which they would recommend in favor of an intervention and below which they would recommend against it.

authors, like guideline developers, need to label the value judgments involved, and acknowledge that judgments may ultimately prove necessary. Box 2 provides an example illustrating how to apply CI approach as the primary criterion for the judgement of imprecision in systematic reviews. For clarity's sake, our discussion in this paper assumes there are no serious concerns regarding the other four GRADE certainty of evidence domains (i.e., risk of bias, inconsistency, indirectness, publication bias).

Box 2 An example of applying confidence interval approach as the primary criterion for the judgement of imprecision in systematic reviews

Consider a systematic review of corticosteroids vs. no corticosteroids for patients with sepsis [3]. A meta-analysis of randomized controlled trials reports that corticosteroids yielded 2.2 fewer deaths per 100 patients with a confidence interval (CI) from 4.1 fewer to 0 fewer (Fig. 1) [3]. Considering the importance of the outcome and using a minimally contextualized approach, the authors could have set an minimally important difference (MID; i.e., the threshold of interest) at a reduction of 0.5 deaths per 100 patients (i.e., 5 deaths per 1,000 patients).

Because the point estimate falls above the MID, authors would rate their certainty that corticosteroids result in an important reduction in death (i.e., the target of certainty rating) [4]. Because the CI crosses the MID of 0.5% (i.e., the effect of corticosteroids might be trivial), authors would rate down at least one level for imprecision. Should the authors rate down two levels for imprecision? The answer is probably not. First, the extent to which the CI crosses the threshold is relatively modest—0.5 per 100. Second, the CI does not include an increase in deaths with corticosteroids. A misguided conclusion of benefit would not, therefore, put patients and clinicians at risk of administration of a lethal intervention. The authors would therefore conclude that corticosteroids probably result in an important reduction in death for patients with sepsis.

3. Based on confidence interval approach GRADE now suggests more frequent rating down two levels for imprecision

An important development in GRADE in the last half decade has been the suggestion that summary of finding tables include plain language summaries of the results [5], and guidance regarding the language to use in such summaries (Box 3) [6]. Considering such plain language summaries enable the GRADE users to consider in the explanation of certainty of evidence whether rating down one, two, or three levels for imprecision makes most sense, and hence provides insight into imprecision rating (Box 3).

3.1. Example 1 [an effect of important benefit, with possibility of important harm]

Returning to the systematic review of corticosteroids for patients with sepsis [3], a meta-analysis of randomized controlled trials (RCTs) reports that on a short term (28–31 days) corticosteroids yields 1.8 fewer deaths per 100 patients with a CI from 4.1 fewer to 0.8 more (Fig. 2) [3]. Rating certainty in relation to the minimally important difference (MID) for benefit (i.e., a reduction of 0.5 deaths per 100 patients), because the point estimate falls above the threshold, authors would rate the certainty that corticosteroids have an important reduction in death (i.e., the target of certainty rating).

Because the CI simultaneously includes important benefit and important harm, the authors would rate down for imprecision. If the authors only rate down one level for imprecision, they would conclude that corticosteroids ''probably'' have an important benefit—a conclusion that is inconsistent with the remaining possibility of important harm. That corticosteroids ''may'' have an important reduction in death is clearly more appropriate, and requires rating down two levels for imprecision.

This situation illustrates the first circumstance in Box 4 that provides principles underlying circumstances in which authors should consider rating down two levels for

Box 3 GRADE plain language summaries and their use in making judgement of imprecision

When high certainty evidence exists, authors of systematic reviews and guidelines can summarize, "effects present"; when moderate certainty evidence exists, authors can conclude, "effects probably or likely present"; and when low quality evidence exists, the plain language summary is "effects possibly present." When the certainty of evidence is very low, authors can make a statement indicating that the evidence is very uncertain [6].

Authors may be appropriately uncomfortable with a plain language summary resulting from rating down only one level for imprecision (e.g., an intervention "probably" has an important effect) when, for instance, a CI includes appreciable harm. In such instances, they may be more comfortable with a summary resulting from rating down two levels for imprecision (i.e., an intervention "may" have an important effect) or a summary resulting from rating down three levels for imprecision (i.e., the evidence is very uncertain about the effect of the intervention).

imprecision. We will illustrate such circumstances using additional examples.

3.2. *Example 2 [an effect of important harm, with possibility of important benefit]*

Consider a systematic review of transcatheter aortic valve implantation (TAVI) vs. surgical aortic valve replacement in patients with severe aortic stenosis. The review of RCTs reports that transapical TAVI results in 5.7 more deaths per 100 patients with a CI from 1.6 fewer to 15.3 more deaths (Fig. 3) [7]. Had the authors considered that the point estimate represents an important harm, without specifying an exact threshold of important harm, they would rate their certainty that transapical TAVI results in an important increase in deaths (i.e., the target of certainty rating). Had they considered that the lower boundary of the CI (1.6 fewer in 100) indicates an important benefit, the authors would conclude that although the point estimate suggests important harm, important benefit remains plausible; thus, they would rate down two levels for imprecision and conclude that transapical TAVI may have an important increase in death (Box 4, circumstance 2).

As shown in this example, authors might not necessarily specify an exact threshold of an important effect. They often find it considerably easier to say whether the point estimate or the end of the CI represents an important effect or not (which also involves value judgment), and that can decide the imprecision judgement.

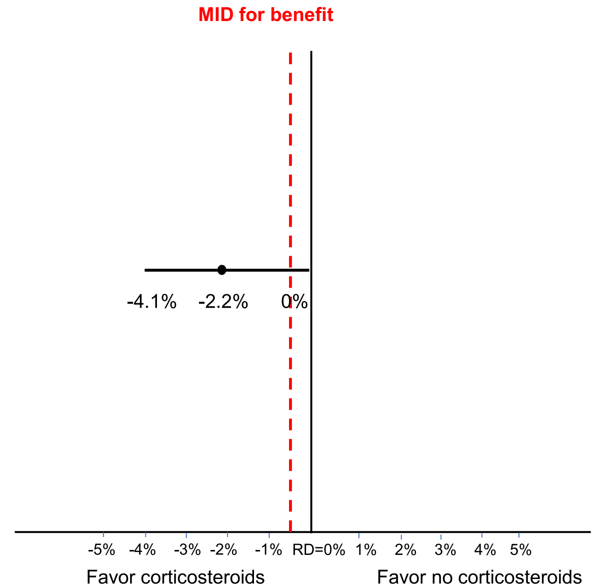

Fig. 1. Effects of corticosteroids vs. no corticosteroids in death for patients with sepsis. *Abbreviations*: MID, minimally important difference.

We will again illustrate how this approach works in Example 6 and Example 7.

Footnote: In Example 2, had the authors not specify an exact threshold of important effect, they could make judgement of whether the end of CI that most favors transapical TAVI indicates an important benefit.

3.3. *Example 3 [trivial effect, with possibility of important benefit and important harm]*

Consider a systematic review of thyroid hormone treatment vs. no treatment in patients with subclinical hypothyroidism. The systematic review reports that thyroid

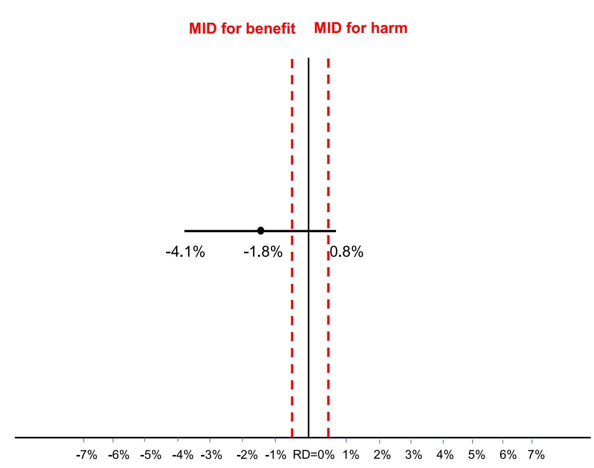

Fig. 2. Effect of corticosteroids vs. no corticosteroids in death at short term for patients with sepsis. *Abbreviations*: MID, minimally important difference.

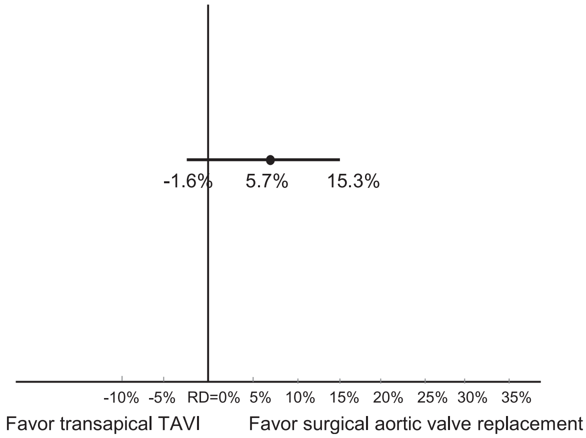

Fig. 3. Effect of transapical TAVI vs. surgical aortic valve replacement in mortality for patients with severe aortic stenosis. *Abbreviations*: TAVI, transcatheter aortic valve implantation.

hormone treatment results in 0.6 fewer cardiovascular events per 100 patients, with a CI from 2.8 fewer to 3.6 more (Fig. 4) [8]. Had the authors set an MID for an important difference in cardiovascular events at 1.5 per 100, the point estimate would fall within the range of trivial effect. The authors would thus rate their certainty that thyroid hormone treatment has a trivial or no effect on cardiovascular events (i.e., the target of certainty rating). As the CI includes both important benefit and important harm, the authors would rate down two levels for imprecision and conclude that thyroid hormone treatment may have trivial or no effect on cardiovascular events (Box 4, circumstance 3).

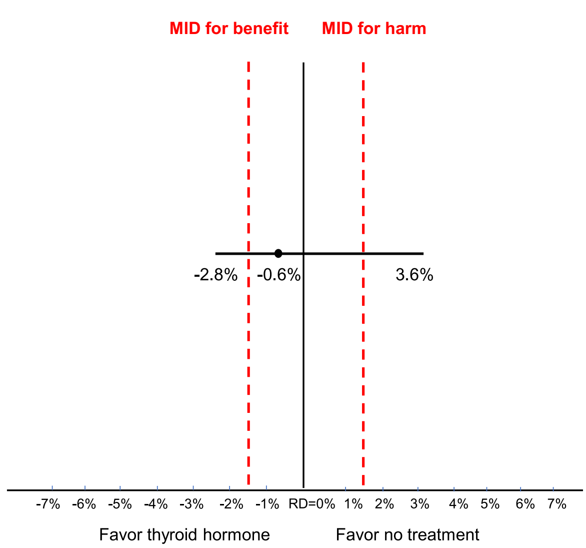

Fig. 4. Effect of thyroid hormone treatment vs. no treatment in cardiovascular event for patients with subclinical hypothyroidism. *Abbreviations*: MID, minimally important difference.

Box 4 Circumstances when one should consider rating down two levels for imprecision based on confidence interval approach using a minimally contextualized approach

Rating certainty in relation to an minimally important difference threshold when point estimate suggests an important effect (circumstance 1–2)

1. When rating the certainty that there is a true important benefit, the point estimate reflects an important benefit, and the boundary of the CI least favorable to the intervention includes the possibility of harm, particularly important harm (Example 1).
2. When rating the certainty that there is a true important harm, the point estimate reflects an important harm, and the boundary of the CI most favorable to the intervention includes the possibility of benefit, particularly important benefit (Example 2).

Rating certainty in relation to minimally important difference thresholds when point estimate suggests a trivial effect (circumstance 3–5)

3. When rating the certainty that the true effect is a trivial or no effect, the point estimate is consistent with a trivial effect, and the CI includes the possibility of both important benefit and important harm (Example 3).
4. When rating the certainty that the true effect is a trivial or no effect, the point estimate is consistent with a trivial effect, and the CI includes the possibility of substantial (possibly large) important harm (Example 4).
5. When rating the certainty that the true effect is a trivial or no effect, the point estimate is consistent with a trivial effect, and the CI includes the possibility of substantial (possibly large) important benefit. (Example 5).

Rating certainty in relation to the null effect threshold (circumstance 6–7)

6. When rating the certainty of nonzero benefit, the point estimate suggests benefit, and the CI includes the possibility of important harm (Example 6).
7. When rating the certainty of nonzero harm, the point estimate suggests harm, and the CI includes the possibility of important benefit (Example 7).

3.4. Example 4 [trivial effect, with possibility of substantial important harm]

We now return to the systematic review of corticosteroids in patients with sepsis. The meta-analysis reports that corticosteroids yield 0.5 more strokes in 100 patients with a CI from 0.3 fewer to 4.3 more (Fig. 5) [3]. Had the authors

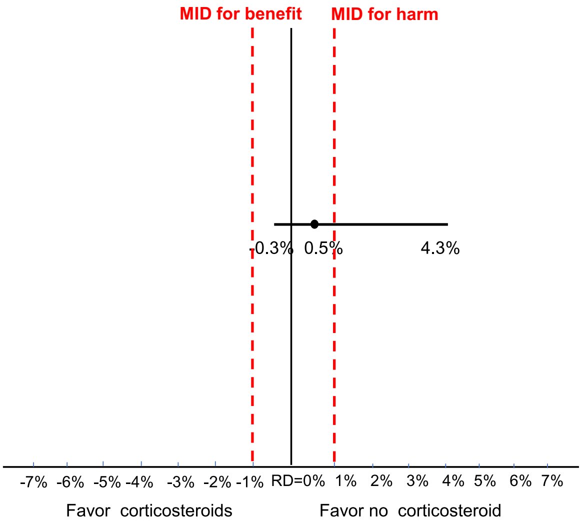

Fig. 5. Effect of corticosteroids vs. no corticosteroids in stroke for patients with sepsis. *Abbreviations*: MID, minimally important difference.

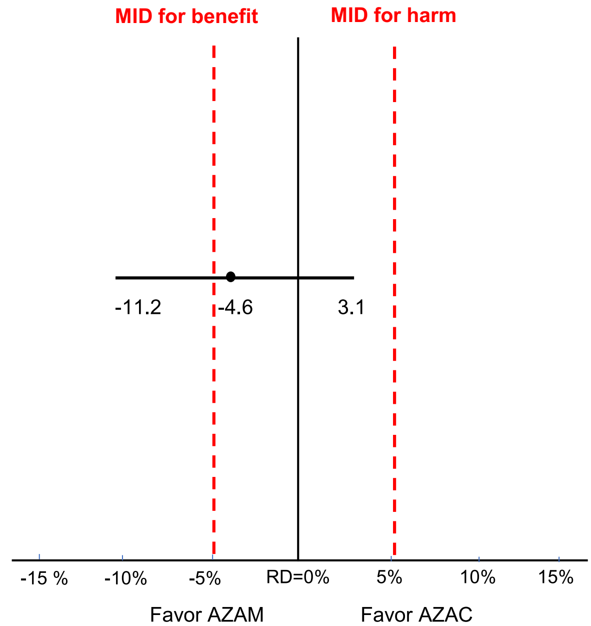

Fig. 6. Effect of azacitidine monotherapy (AZAM) vs. azacitidine combination (AZAC) in thrombocytopenia for patients with acute myeloid leukemia. *Abbreviations*: MID, minimally important difference.

set an MID for benefit or harm at a difference of one stroke per 100, the point estimate would fall within the range of trivial or no effect. The authors would rate their certainty that corticosteroids have trivial or no effect on stroke (i.e., the target of certainty rating). Because the CI crosses the boundary of an important effect, authors would rate down at least one level for imprecision.

The question then arises whether the boundary of CI suggesting maximum harm, 4.3% increase in stroke, is large enough to warrant rating down two levels for imprecision. Authors might think of this issue in several ways. First, how much patients value the outcome — the more important the outcome, the greater the likelihood of rating down two levels. Authors might also consider judgement of the magnitude of such an increase: Would it represent a moderate or large effect? The larger the magnitude, the more likely authors would rate down two levels. Another approach would be to consider the plain language statements that would accompany decisions regarding whether to rate down one or two levels. Would one be comfortable with a statement that it is ‘‘likely’’ that corticosteroids have trivial or no effect on stroke when it remains possible that corticosteroids result in a 4.3% increase. If not — and the more comfortable statement would be that they ‘‘may’’ result in little or no effect on stroke — rating down two levels would be more appropriate. Our judgment would be stroke is sufficiently important and 4.3% represents a sufficiently large effect that we would rate down two levels for imprecision (Box 4, circumstance 4).

3.5. Example 5 [trivial effect, with possibility of substantial important benefit]

Consider a systematic review of RCTs including patients with acute myeloid leukemia that compared azacitidine monotherapy (AZAM) vs. azacitidine combination. The meta-analysis shows that AZAM yields, in 100 patients, 4.6 fewer with thrombocytopenia, with a CI from 11.2 fewer to 3.1 more (Fig. 6) [9]. Had the authors set an MID for benefit or harm at a difference of five thrombocytopenia events per 100, the point estimate would fall within the range of trivial or no effect. The authors would rate their certainty that AZAM, in comparison to azacitidine combination, has a trivial or no effect on thrombocytopenia (i.e., the target of certainty rating). Because the CI crosses the MID for benefit, authors would certainly rate down at least one level for imprecision.

If the authors considered that a reduction in thrombocytopenia of 11.2% was sufficiently large and was more comfortable with a statement that AZAM ‘‘may’’ (rather than ‘‘probably’’) result in trivial or no effect on thrombocytopenia, the authors would rate down two levels for imprecision (Box 4, circumstance 5).

3.6. Example 6 [an effect of benefit, with possibility of important harm]

Consider a systematic review including one RCT with patients with vasculitis that compares a reduced-dose regimen of glucocorticoids with a standard-dose regimen. The systematic review reports that the reduced-dose regimen yields 2.1 fewer deaths per 100 patients, with a CI from 6 fewer to 3.6 more (Fig. 7) [10]. Had the authors rated their certainty in relation to the null effect (i.e., risk difference = 0), they would rate their certainty that the reduced-dose regimen of glucocorticoids reduces mortality (i.e., the target of certainty rating). Had the authors considered that the upper boundary of the CI (i.e., 3.6 more in

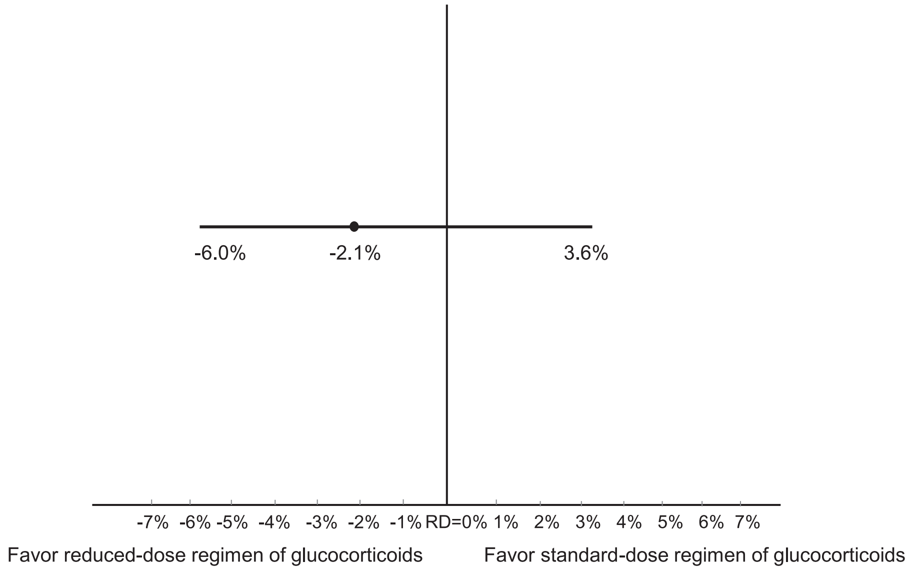

Fig. 7. Effect of reduced-dose regimen vs. standard-dose regimen of glucocorticoids in mortality for patients with vasculitis.

100) indicates an important harm, the authors would conclude that although the point estimate suggests a benefit, important harm remains plausible.

Using the GRADE plain language statement, the authors could consider whether they would be comfortable with a statement that it is ‘‘likely’’ that, compared with a standard-dose regimen, a reduced-dose regimen has a benefit in reducing death when it remains possible that the reduced-dose regimen has an important increase in death. If not—and the more comfortable statement would be that the reduced-dose regimen ‘‘may’’ result in reducing death—they would rate down two levels for imprecision (Box 4, circumstance 6).

Footnote: In Example 6, had the authors not specify an exact threshold of important harm, they could make judgement of whether the end of CI that least favors reduced-dose regimen of glucocorticoids indicates an important harm.

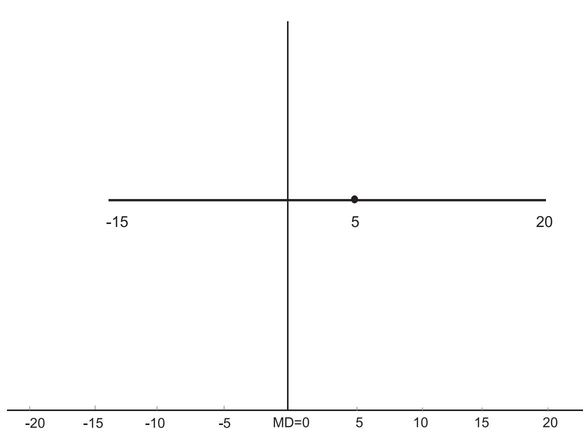

Fig. 8. Effect of low intensity pulsed ultrasound (LIPUS) vs. no ultrasound in return to work for patients with any type of fracture.

3.7. Example 7 [an effect of harm, with possibility of important benefit]

Consider a systematic review of low intensity pulsed ultrasound (LIPUS) vs. no ultrasound in patients with any type of fracture [11]. A meta-analysis of RCTs reports that LIPUS results in five additional days before patients could return to work, with a CI from 15 fewer to 20 more days (Fig. 8). Rating certainty in relation to the null effect (i.e., MD = 0), the authors would rate their certainty that LIPUS increases the days off required before returning to work (i.e., the target of certainty rating). Without specifying a precise threshold of MID for benefit, the authors could have considered that the boundary of the CI most favorable to LIPUS (15 days earlier) indicates an important benefit, and thus would rate down two levels for imprecision (Box 4, circumstance 7).

As shown in Example 6 and Example 7, when rating certainty in a nonzero benefit (or a nonzero harm) and considering rating down two levels for imprecision, authors need to think about whether the boundary of the CI least (or most) favorable to the intervention includes the possibility of important harm (or important benefit). Authors often struggle with specifying a threshold of important difference. As illustrated in these two examples, authors might find it considerably easier to say whether the end of the CI represents an important harm (or an important benefit). That judgement can then determine the imprecision rating.

Example 1 to Example 7 illustrated circumstances in which CIs are considerably crossing the threshold(s) of interests that one or both boundaries of CIs suggest inferences appreciably different from point estimate. In these circumstances GRADE suggests considering rating down two levels for imprecision.

Box 5 Circumstances when one should consider rating down two levels for imprecision based on optimal information size calculation using a minimally contextualized approach

1. For dichotomous outcomes, when the ratio of the upper to the lower boundary of the CI is more than 2.5 for odds ratio or three for risk ratioR (Appendix A, Example 1).
2. For continuous outcomes, when the total sample size of a meta-analysis is smaller than 30–50% of the OIS (Appendix A, Example 2).

4. When the confidence interval does not cross threshold(s) of interest and the relative effect is large, GRADE suggests considering whether the optimal information size is met

4.1. When systematic review and guideline authors should check optimal information size

Considering OIS refers to considering whether the total number of participants or events in the meta-analysis is more than the number of participants or events generated by a conventional sample size calculation for a single adequately powered trial [1,2]. When the CI overlaps with threshold(s) of interest, authors would rate down for imprecision and do not need to consider the OIS.

When the CI does not overlap with the threshold(s) of interest and the effect is sufficiently large and [e.g., relative risk (RR) reduction or an RR increase over 30%] that the authors consider the results implausible, authors should consider implementing the OIS [1,2,12]. If the OIS is met, authors do not need to rate down for imprecision; otherwise, authors should rate down. It is important to notice that because the calculation of OIS is based on relative estimates of effect [1], all assessments and steps in OIS approach focus on the relative estimate of effect (as opposed to the CI approach, which is done based on absolute estimates of effect).

4.2. Based on optimal information size calculation when GRADE suggests rating down two levels for imprecision

Simulations conducted to inform GRADE guidance for addressing imprecision of dichotomous outcomes in the context of network meta-analysis provided insights into how many levels to rate down for imprecision in pairwise meta-analysis when considering the OIS [12]. These simulations suggested that when the ratio of the upper to the lower boundary of the CI is higher than 2.5 for odds ratios and three for RRs, the sample size is, for any reasonable combination of baseline risk and treatment effect, very far from meeting the OIS. Therefore, authors would not need to calculate the OIS and can rate down the certainty of evidence by two levels (Box 5, circumstance 1). If, on the other hand, the effect is large and the ratio is less than these thresholds, authors should calculate the OIS and compare it to the sample size available in the meta-analysis. If, in this latter situation, the OIS criteria is not met, authors would rate down the certainty of the evidence by one level for imprecision. The first example in Appendix A illustrates how to apply the OIS approach for rating imprecision of dichotomous outcomes.

Using the OIS in the context of continuous outcomes presents complexities that do now allow creating guidance based on the boundaries of the CI. If authors are not, however, confident of the MID or the standard deviation (SD) needed for the calculation of OIS, they could use an effect size of 0.2 SDs that represents a small effect [13]. This results in a total sample size of approximately 800 (400 per group) [1]. GRADE suggests, based on OIS approach, rating down two levels for imprecision if the total sample size of meta-analysis is smaller than arbitrary threshold of 30–50% of OIS (Box 5, circumstance 2). If the authors choose to be more conservative, they could choose 50% of OIS as a threshold (i.e., 400 overall); if they choose to be less conservative, they could use 30% of OIS as the threshold (i.e., 240 overall). The second example in Appendix A illustrates how to apply the OIS approach for rating imprecision of continuous outcomes.

4.3. When the baseline risk is very low, GRADE suggests being more restrained in rating down for imprecision

The third example in Appendix A illustrates the circumstances in which failure to meet the OIS will not necessarily require authors to rate down the certainty of evidence.

5. Based on confidence interval approach when GRADE suggests rating down three levels for imprecision

The GRADE guidance of how to choose target of certainty rating identified circumstances in which CIs are so wide (i.e., the two boundaries of CI suggest very different inferences) that authors are very uncertain regarding the true effect and authors do not need to determine the target of certainty rating [4]. In these circumstances, one can rate down three levels for imprecision. Exactly how wide a CI should be when one can make such decision is a matter of value judgement.

For instance, in Example 6 and Example 7, had the authors considered both boundaries of the CI presented large effects (i.e., the CI includes both large benefit and large harm), they would consider rating down three levels for imprecision, and would conclude that they were uncertain about the effects of interventions.

A companion paper addressing precision ratings in partially and fully contextualized settings includes a major focus on rating down by three levels for imprecision, and for further guidance authors can apply the suggestions there to the minimally contextualized approach [14]. GRADE users should also stay alert to the possibility of double counting imprecision and inconsistency when using random-effects models to conduct their meta-analysis. For further guidance, authors can also refer to the companion paper [14].

6. Discussion

This article emphasizes the usefulness of CI approach for imprecision rating in systematic reviews, reserving OIS approach to situations of implausibly large treatment effects. We encourage authors using minimally contextualized approaches to be alert to the possibility of rating down more than one level for imprecision. When the issue arises, authors are likely to find reference to Boxes 4 and 5, and the associated examples, helpful.

Acknowledgments

The authors would like to thank the author of one of the systematic reviews we used as examples in this paper: Luis Enrique Colunga Lozano. The authors also would like to thank GRADE Working Group members who gave valuable comments and suggestions on the draft of this article. The authors also thank attendants of GRADE Working Group meetings who have contributed to the article during group discussions.

Appendix A

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.07.014.

References

[1] Guyatt GH, 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;64:1283–93.

[2] Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol 2017;87:4–13.

[3] Rochwerg B, Oczkowski SJ, Siemieniuk RAC, Agoritsas T, Belley-Cote E, D'Aragon F, et al. Corticosteroids in sepsis: an updated systematic review and meta-analysis. Crit Care Med 2018;46:1411–20.

[4] Zeng L, Brignardello-Petersen R, Hultcrantz M, Siemieniuk RAC, Santesso N, Traversy G, et al. GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings. J Clin Epidemiol 2021;137:163–75.

[5] Carrasco-Labra A, Brignardello-Petersen R, Santesso N, Neumann I, Mustafa RA, Mbuagbaw L, et al. Improving GRADE evidence tables part 1: a randomized trial shows improved understanding of content in summary of findings tables with a new format. J Clin Epidemiol 2016;74:7–18.

[6] Santesso N, Glenton C, Dahm P, Garner P, Akl EA, Alper B, et al. GRADE guidelines 26: informative statements to communicate the findings of systematic reviews of interventions. J Clin Epidemiol 2020;119:126–35.

[7] Siemieniuk RA, Agoritsas T, Manja V, Devji T, Chang Y, Bala MM, et al. Transcatheter versus surgical aortic valve replacement in patients with severe aortic stenosis at low and intermediate risk: systematic review and meta-analysis. BMJ 2016;354:i5130.

[8] Feller M, Snel M, Moutzouri E, Bauer DC, de Montmollin M, Aujesky D, et al. Association of thyroid hormone therapy with quality of life and thyroid-related symptoms in patients with subclinical hypothyroidism: a systematic review and meta-analysis. JAMA 2018;320:1349–59.

[9] Lozano LEC, Nampo FK, Agarwal A, Desai P, Litzow M, Sekeres MA, et al. Less intensive antileukemic therapies (monotherapy and/or combination) for older adults with acute myeloid leukemia who are not candidates for intensive antileukemic therapy: a systematic review and meta-analysis. PLoS One 2011;17:e0263240.

[10] Xiao Y, Guyatt G, Zeng L, Jayne DR, Merkel PA, Siemieniuk RA, et al. The comparative efficacy and safety of alternative glucocorticoids regimens in patients with ANCA-associated vasculitis: a systematic review. BMJ Open 2011;12:e050507.

[11] Schandelmaier S, Kaushal A, Lytvyn L, Heels-Ansdell D, Siemieniuk RAC, Agoritsas T, et al. Low intensity pulsed ultrasound for bone healing: systematic review of randomized controlled trials. BMJ 2017;356:j656.

[12] Brignardello-Petersen R, Guyatt GH, Mustafa RA, Chu DK, Hultcrantz M, Schünemann HJ, et al. GRADE guidelines 33: addressing imprecision in a network meta-analysis. J Clin Epidemiol 2021;139:49–56.

[13] Cohen J. A power primer. Psychol Bull 1992;112:155–9.

[14] Schünemann HJ, Neumann I, Hultcrantz M, Zeng L, Murad MH, Izcovich A, et al. GRADE Guidance article: rating down for imprecision in the partially or fully contextualized approaches to assessing certainty of evidence. J Clin Epidemiol 2022. Online ahead of print.