Core GRADE 2: Choosing the Target of Certainty Rating and Assessing Imprecision

Gordon Guyatt, distinguished professor [1] [2] [3] §, Linan Zeng, professor [3] [4] §, Romina Brignardello-Petersen, associate professor [1], Manya Prasad, associate professor [5], Hans De Beer, methodologist [6], M Hassan Murad, professor [7], Alfonso Iorio, professor [1] [2], Arnav Agarwal, methodologist [1] [2] [3], Liang Yao, assistant professor [8], Thomas Agoritsas, associate professor [1] [3] [9], Jamie Rylance, associate professor [10], Reem A Mustafa, professor [1] [11], Per Olav Vandvik, professor [1] [12] [13], Prashanti Eachempati, adjunct Professor [3] [14] [15], Chunjuan Zhai, associate professor [16], Lingli Zhang, professor [3] [4], Victor Montori, professor [17] [18], Monica Hultcrantz, head of HTA Region Stockholm [19] [20]

1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
2. Department of Medicine, McMaster University, Hamilton, Ontario, Canada
3. MAGIC Evidence Ecosystem Foundation, Oslo, Norway
4. Pharmacy Department/Evidence-based Pharmacy Centre/Children's Medicine Key Laboratory of Sichuan Province, West China Second University Hospital, Sichuan University; Sichuan University and Key Laboratory of Birth Defects and Related Disease of Women and Children, Ministry of Education; West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China
5. Clinical Research and Epidemiology, Institute of Liver and Biliary Sciences, New Delhi, India.
6. Guide2Guidance, Lemelerberg 7, 3524 LC Utrecht, the Netherlands.
7. Evidence-based Practice Center, Mayo Clinic, Rochester, MN 55905, USA
8. Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore
9. Division General Internal Medicine, University Hospitals of Geneva, University of Geneva, Geneva, Switzerland
10. Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, United Kingdom
11. Department of Medicine, University of Kansas Medical Center, Kansas City, MO, USA
12. Institute of Health and Society, University of Oslo Faculty of Medicine, Oslo, Norway
13. Department of Medicine, Lovisenberg Diakonale Hospital, Oslo, Norway
14. Peninsula Dental School, University of Plymouth, United Kingdom
15. Faculty of Dentistry, Manipal University College Malaysia
16. Department of Cardiology, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong 250021, China
17. Division of Endocrinology, Department of Medicine, Mayo Clinic, Rochester, MN, USA
18. Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN, USA
19. HTA Region Stockholm, Centre for Health Economics, Informatics and Health Care Research (CHIS), Stockholm Health Care Services, Sweden
20. Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

§ Gordon Guyatt and Linan Zeng contributed equally and are joint first authors.

Corresponding Author: Gordon Guyatt, guyatt@mcmaster.ca; Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada; Department of Medicine, McMaster University, Hamilton, ON L8S 4L8, Ontario, Canada

Abstract

This second article in a seven-part series presents the Core GRADE approach to deciding on the target of the certainty rating - what it is in which we are rating certainty - and whether and how much to rate down the certainty of evidence due to imprecision.

Core GRADE users assess if an effect is important or not in relation to the minimal important difference (MID). Alternatively, they can rate certainty regarding whether there is a true underlying treatment effect, positive or negative which means relating to the null threshold (relative risk of 1.0, risk difference of 0 or mean difference of 0).

The location of the point estimate of effect in relation to the chosen threshold determines the target. Using MID thresholds, a point estimate greater than the MID suggests an "important effect"; less than the MID, an "unimportant" or "little to no effect". Users then rate down for imprecision if the 95% confidence interval (CI) crosses the MID for benefit or harm.

When users assess certainty in a true underlying effect, they rate down for imprecision if the 95% CI crosses the null and thus includes both benefit and harm. If the point estimate is very near the null, users rate their certainty in "little to no effect" (i.e. an unimportant effect).

If the CI includes both an important benefit and an important harm, Core GRADE users will typically rate down two levels for imprecision. Finally, for either threshold, if the CI does not cross a threshold but the effect is very large, Core GRADE users will consider rating down for imprecision if the sample size or number of events do not meet a threshold based on sample size calculations for a single RCT – an approach we refer to as the optimal information size.

Curricular Objective

After reading this paper, Core GRADE users will be able to:
- Make choices about whether to assess their certainty in an important or unimportant effect, or assess their certainty in a true underlying treatment effect.
- Make judgements about rating down certainty of evidence once or twice for imprecision by considering the width of the confidence interval and the appropriate plain language summary.
- Decide when and how to consider the optimal information size in making judgements about rating down certainty of evidence for imprecision.

Take Home Messages
- Core GRADE users may assess their certainty in an important or unimportant effect (using the minimal important difference threshold), which often simplifies and clarifies ratings of imprecision and interpretation of results.
- In some situations, when wanting to minimize value judgments, Core GRADE users may use the null threshold to assess their certainty in a true underlying treatment effect.
- When deciding whether to rate down certainty for imprecision, Core GRADE users will consider whether the CI crosses the chosen threshold and will usually rate down two levels if the CI includes both an important benefit and an important harm.
- If the CI does not cross a threshold but the effect is very large, Core GRADE users will consider rating down for imprecision if the sample size or number of events are too small (not meeting the optimal information size).

1. Introduction

This is the second in a series of papers describing the essentials of GRADE in rating certainty of evidence and grading recommendations, focusing on evidence addressing alternative care options from the perspective of the patient. In the first paper, we described our overall Core GRADE approach noting that GRADE's certainty rating represents our confidence that the true effect lies on one side of a threshold (such as an important difference) or in a particular range (such as an unimportant or little to no effect). We further noted that randomized controlled trials (RCTs) start as high certainty evidence in GRADE's four-category certainty approach (high, moderate, low, and very low) and non-randomized studies of interventions (NRSI, observational) start at low certainty.[1]

This paper includes two parts. We start with a key element of the GRADE approach: choosing the target of our certainty rating – that is, what it is in which we rate our certainty[2][3]. Secondly, we provide guidance regarding decisions about rating down for imprecision – one of the 5 domains that Core GRADE users consider when assessing the certainty of evidence in both RCT and NRSI[4]. The following papers in this series will deal with the other four domains.

2. In what should we rate our certainty?

In assessing the effect of an intervention, our primary interest is whether it outperforms alternatives, such as standard care or other existing treatments. If there is no difference in benefit outcomes, a guideline panel will unlikely recommend and clinicians are unlikely to use the new treatment unless it offers other advantages such as reduced harms or burdens. Moreover, merely identifying the presence of an effect is frequently insufficient to recommend a treatment: we also need to know whether the effect is large enough to be important for patients. The question "Is there any effect compared to the alternative?" corresponds to using the threshold of "null effect", while "Is the effect important?" aligns with using the minimal important difference (MID). The MID, a crucial concept in clinical studies and GRADE methodology, represents the smallest change in a single outcome that patients perceive as important.[5]

The focus of Core GRADE is on these two questions of 1) whether there is an effect compared to the alternative (i.e. using the null threshold) and 2) whether the effect is large enough to be important for patients (i.e. using the MID) (Figure 1). Use of additional thresholds of moderate and large effects has proved challenging for GRADE users and in our judgement does not provide important incremental value in making sound and optimally useful ratings of certainty. Nevertheless, Appendix 1 presents judgements of small, moderate and large effects that Core GRADE users may sometimes want to consider but are not part of Core GRADE.[6]

Figure 1. Thresholds and ranges for rating certainty of evidence. Besides the Core thresholds of null effect and MID, one may consider two other thresholds, the moderate effect threshold that demarcates small versus moderate effects, and the large effect threshold that demarcates moderate versus large effects.
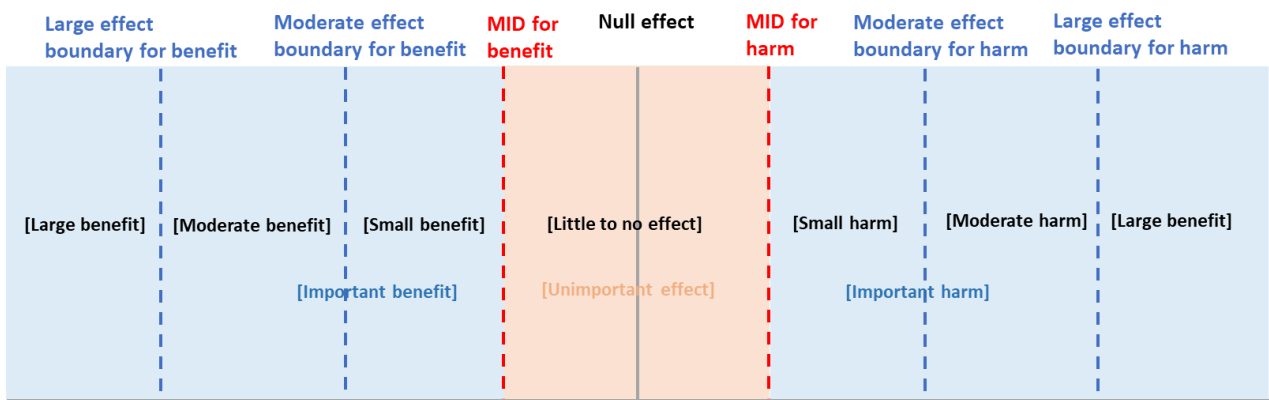
Deciding what it is in which we are rating our certainty requires three steps (illustrated in Figure 2). First Core GRADE users choose if they are interested in whether an effect is important or not or whether there is a true underlying effect compared to the alternative.

In the next step Core GRADE users establish the effect estimates through meta-analysis. An important choice Core GRADE users will face is whether, in their analysis, to use fixed effect or random effect statistical models. Appendix 2 provides key issues about the approaches.

In rating certainty of evidence Core GRADE users typically consider absolute rather than relative effects. For binary outcomes, Core GRADE users obtain the best estimate of the risk difference and its 95% CI by applying the relative risk to an estimate of the baseline risk. Appendix 3 illustrates how this is done. A subsequent paper in this series focusing on Summary of Findings tables presents additional information regarding absolute effects, and how to present continuous outcomes. [7]

Finally, Core GRADE users assess the magnitude of the absolute effect estimate in relation to the chosen threshold. The process for these steps is addressed in the subsequent discussion and illustrated in Figure 2.

Figure 2. Steps for deciding target of certainty rating

MID: Minimal Important Difference; WMD: Weighted Mean Difference; RD: Risk Difference
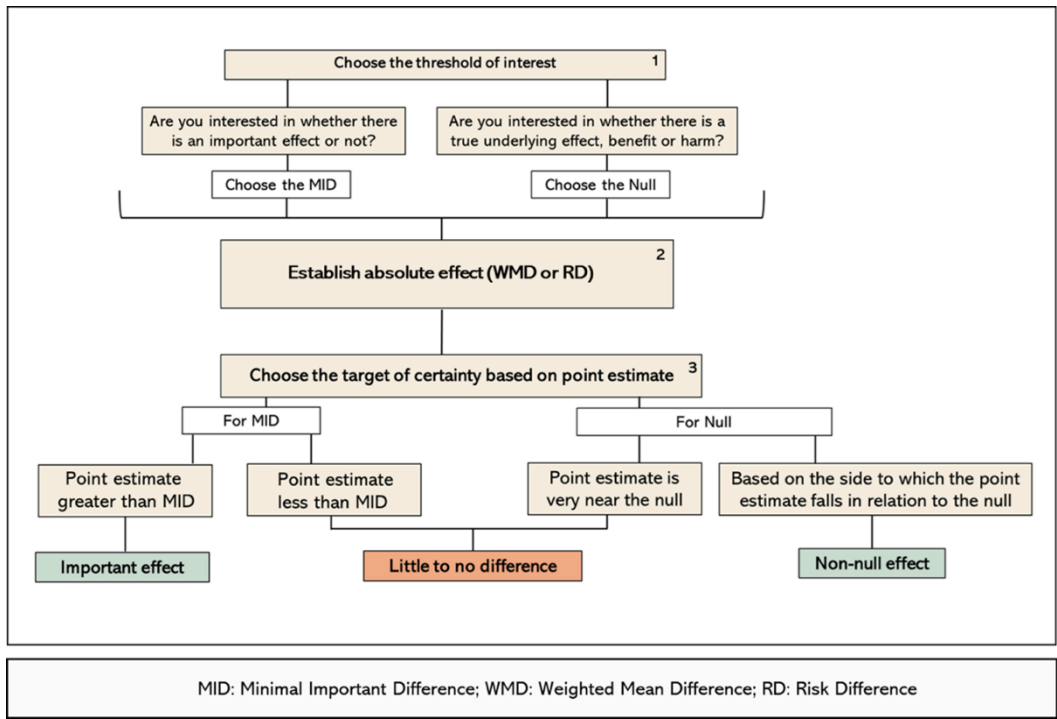
2.1 Assessing if the effect is important or not (i.e. using the MID threshold)
When considering whether an effect is important, Core GRADE users must focus on absolute (i.e., risk differences) rather than relative effects. The reason is, as we explained in the first article in this series, it is absolute rather than relative effects that are important to patients. This is because a 50% relative risk reduction (risk ratio of 0.5) could represent a 1% absolute reduction (from a baseline risk of 2% in control group to 1% in intervention group) or a much larger 20% absolute reduction (from 40% to 20%).

If Core GRADE users are interested in whether an effect is important, they will thus need to make a value judgment regarding the importance of the outcome and, in particular, the threshold that delineates an important from an unimportant effect, i.e. the MID. The values and preferences that drive this choice should be those of the patients or other target populations (e.g., the general public).

Guideline development and Health Technology Assessment (HTA) require judgments regarding how people value the benefits, harms, and burdens of the interventions under consideration. Specifying MIDs, either using established MIDs (most likely to be available for patient-reported outcomes such as pain, functional status, or quality of life [5 8]) or their own estimates (generated from, for example, existing literature or their clinical experience [9]) has proven helpful in facilitating the trade-offs between desirable and undesirable consequences of interventions. In the seventh and final article in this series we address, in some detail, the issues of choosing MIDs for key outcomes.[10]

If Core GRADE users have chosen the MID threshold and the point estimate from the meta-analysis represents an effect greater than the MID ([a] in Figure 3), review authors will rate their certainty that the true effect is an important benefit (or, if favouring the comparator, an important harm). If the point estimate represents an effect less than the MID ([b] in Figure 3), they will rate their certainty in an unimportant (synonym little to no) effect. [3]

Figure 3. Assessing if the effect is important or not. (a) As the point estimate falls above the MID, the target of the certainty rating is that the true effect is important. (b) As the point estimate falls below the MID, the target of the certainty rating is that the true effect is unimportant (little to no effect).
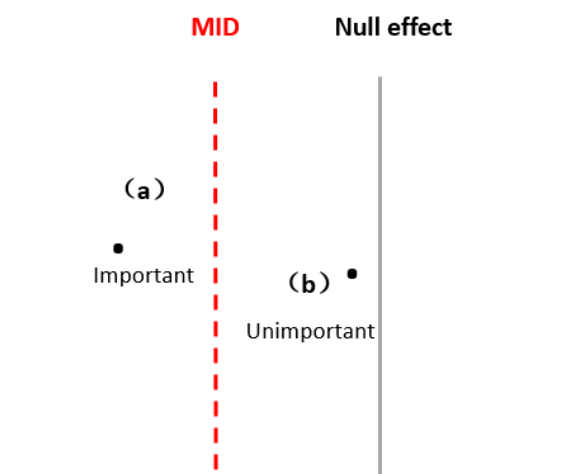
2.2 Assessing whether there is a true underlying treatment effect (i.e. using the null as the threshold)
For several reasons, the null represents an alternative threshold to the MID in systematic reviews. Evidence regarding the distribution of values and preferences in the population of interest is typically limited, making inferences regarding the MID challenging. Further, review authors may not see their mandate as including the search and interpretation of relevant evidence about MIDs. Finally, review authors may wish to leave the value judgments involved in choosing specific MIDs to HTA and guideline practitioners who typically consult a wider group of individuals and often in a structured way.

If Core GRADE users have chosen the null they will, based on where the point estimate falls in relation to the null, typically rate certainty that a true beneficial or a harmful effect exists. If, however, the point estimate is very near the null, because the intuitive inference in such situations is that the true effect represents little to no difference between intervention and control, they will rate their certainty in an unimportant effect (as in [b] in Figure 3). [11] So, while choosing the null usually avoids specifying MIDs, it will not always do so. Appendix 4 provides an expanded discussion of this issue. The difficulty in altogether escaping considerations of importance when choosing the null as a threshold may lead Core GRADE users to prefer the MID as a threshold.

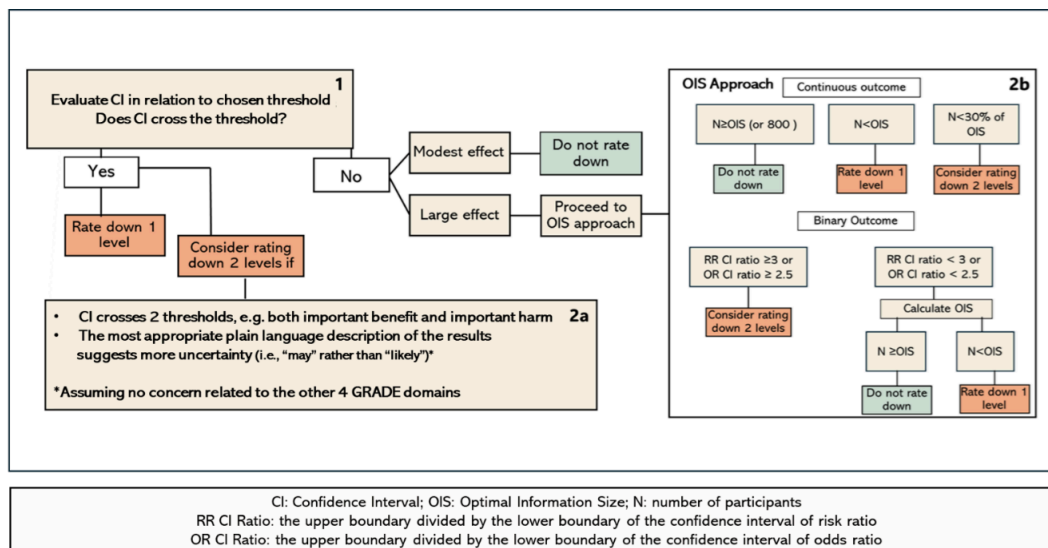## 3. Rating the certainty of evidence: Imprecision

After deciding on the target of certainty rating, Core GRADE users assess whether limitations exist in any one of five GRADE domains (imprecision, inconsistency, risk of bias, indirectness, and publication bias). The following discussion addresses how Core GRADE users can make judgments regarding imprecision.

3.1 What is imprecision?

Studies of interventions seek to estimate the true underlying treatment effect. A meta-analysis provides our best estimate of the effect (the point estimate) and the confidence intervals (CI) provide the bounds in which the true effect plausibly lies. The most commonly used CI is the 95% CI. The width of confidence intervals provides key information regarding the extent of imprecision, thus informing the impact of random error on certainty of evidence.

We will now describe GRADE's approach to making the judgment of whether the CI is sufficiently wide that Core GRADE users should rate down for imprecision, and whether they should rate down one or two levels. The judgments regarding rating down once or twice have proven particularly challenging for Core GRADE users and, in the current discussion, we build on clarification in previous papers. [4][6] Figure 4 presents the steps Core GRADE users take in making these judgments.

Figure 4. Steps for rating imprecision

CI: Confidence Interval; OIS: Optimal Information Size; N: number of participants
RR CI Ratio: the upper boundary divided by the lower boundary of the confidence interval of risk ratio
OR CI Ratio: the upper boundary divided by the lower boundary of the confidence interval of odds ratio

3.2 Rating down (or not) for imprecision

When deciding whether to rate down certainty for imprecision, Core GRADE users will consider whether the CI crosses the chosen threshold. For instance, consider the pooled effect estimate from a hypothetical systematic review of RCTs illustrated in Figure 5. For (a) in Figure 5, whether Core GRADE users are rating certainty with respect to a non-null effect (null being a risk difference [RD] of 0%) or in an important effect (the MID threshold of 1%), the CI does not cross either threshold, and they will not rate down their certainty for imprecision. Assuming they have no concerns about the other four GRADE domains, they will have high certainty that there is a non-null effect, as well as that the effect is important.

For (b) in Figure 5, decisions regarding rating down certainty will differ depending on the threshold. When using the null, as the CI does not cross the threshold, Core GRADE users will not rate down their certainty for imprecision. When using the MID, as the CI crosses the threshold, they will rate down for imprecision.

For (c) in Figure 5, whether Core GRADE users are rating their certainty in relation to the null or the MID, the CI crosses the threshold and Core GRADE users will rate down for imprecision.

Figure 5. Example of how the target of certainty rating (above the MID or above the null) will affect the rating of imprecision. (a) Core GRADE users will not rate down for imprecision in either case. (b) If the target of certainty rating is an important effect (above the MID), Core GRADE users will rate down for imprecision, but if the target of certainty rating is an effect above the null they will not. (c) Core GRADE users will rate down for imprecision in both cases.
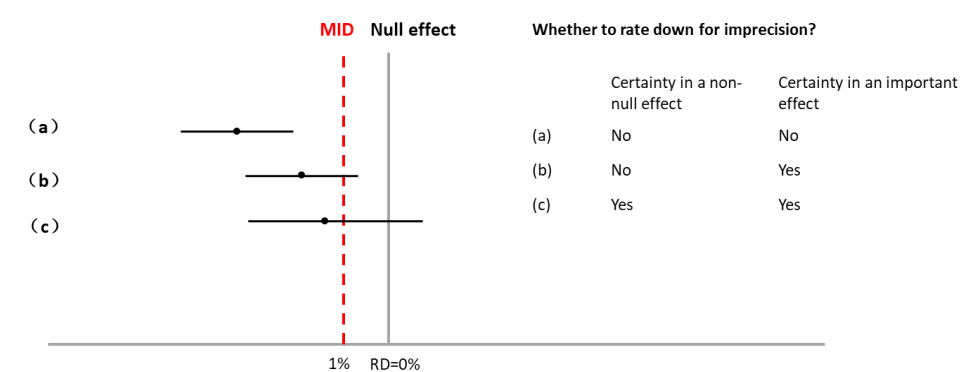
Consider a systematic review of corticosteroids versus no corticosteroids for patients with community-acquired pneumonia (Figure 6).[12] The meta-analysis of RCTs reported that corticosteroids yielded 29 fewer deaths per 1000 patients with a CI from 52 fewer to 6 fewer. If review authors have chosen the null as their threshold, they will rate their certainty in a mortality reduction and will not rate down for imprecision.

If review authors have chosen the MID as their threshold and set their MID at a difference of 10 deaths per 1000 patients (MID1 in Figure 6), because the point estimate is greater than their threshold, they would rate down their certainty in an important mortality reduction. Further, they would rate down for imprecision since the CI crosses the MID threshold. Had they chosen an MID of 5 deaths per 1000 patients (MID2 in Figure 6), they would also rate their certainty that an important effect exists but would not rate down for imprecision because the CI does not cross the MID threshold.

Figure 6. An example of how rating down for imprecision depends on the choice of MID in a systematic review of corticosteroids versus no corticosteroids on mortality in patients with community-acquired pneumonia. If the review authors set the MID at MID1, they will rate down for imprecision. If they set the MID at MID2, they will not rate down for imprecision.
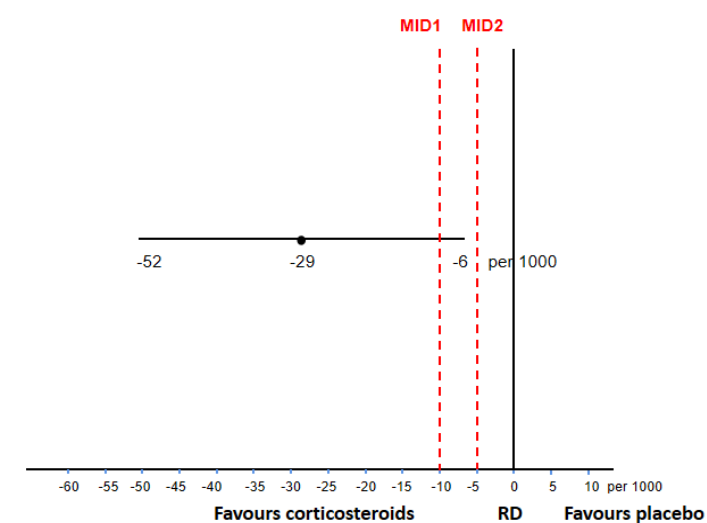
When Core GRADE users have chosen the MID as their threshold and the point estimate is less than the MID, they will rate their certainty that the true treatment effect is unimportant (i.e. little to no effect) (Figure 7 all point estimates). As described in section 2.2, when Core GRADE users have chosen the null as the threshold and the point estimate clearly suggests an unimportant effect (i.e., the point estimate falls very close to the null) they will instead rate certainty in little to no effect by relating the confidence interval to the MID (see Appendix 4 for further illustration). In either case, they will not rate down for imprecision if the CI crosses neither threshold (Figure 7 [a]). If the CI crosses one ([b], [c]), or both thresholds [d] in Figure 7) they will rate down for imprecision.

Figure 7. Rating certainty in little to no effect and rating down for imprecision when the CI crosses the MID.
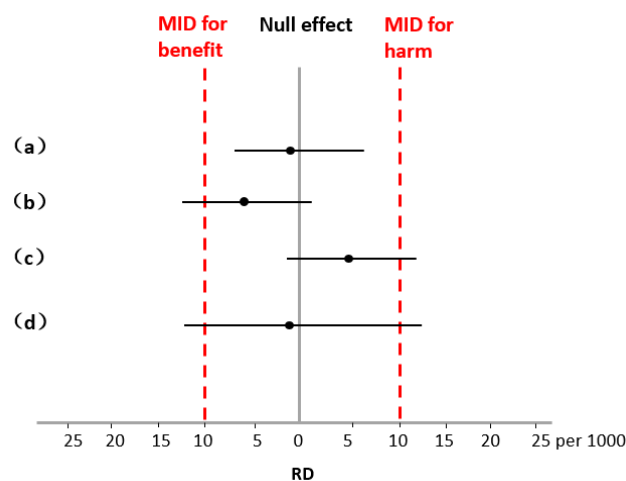
4. Rating down once or twice for imprecision
As the CI gets wider, Core GRADE users become progressively more uncertain whether the truth is consistent with an important or unimportant effect, or whether it reflects a non-null effect. To reflect the degree of uncertainty influenced by imprecision of evidence, Core GRADE users can consider rating down one or two levels for imprecision.

4.1 A possible role for plain language statements in deciding whether to rate down once or twice
Stating results in plain language that both clinicians and patients will easily understand is important in making GRADE optimally useful for clinical practice. GRADE has therefore provided guidance in making such statements (Table 1).[13] We will return to these statements in the sixth paper in this series in which we discuss GRADE Summary of Findings tables; we introduce them here because they can help decide on rating down once or twice for imprecision. [8]

Table 1. GRADE plain language statements when using the null effect or MID thresholds

Certainty	Plain language summary	
	Null effect as the threshold	MID as the threshold
High	"Treatment has a benefit", or "Treatment improves outcome X"	"Treatment has an important benefit, or "Treatment has little to no benefit"
Moderate	"Treatment likely has a benefit", or "Treatment likely improves outcome X"	"Treatment likely has an important benefit", or "Treatment likely has little to no benefit"
Low	"Treatment may have a benefit", or "Treatment may improve outcome X"	"Treatment may have an important benefit", or "Treatment may have little to no benefit"

| Very low | "We are very uncertain about whether treatment has a benefit" |

Note: The plain language summary pertains to both beneficial and harmful outcomes. We chose benefit here for illustration.

4.2 Rating down once or twice for imprecision - general principles

We suggest two useful considerations when deciding whether to rate down twice: first, does the CI cross more than one threshold (e.g., includes both important benefit and important harm). Second, considering the GRADE plain language, is the most appropriate message that a particular effect likely exists or that it may exist.

Consider a systematic review comparing reduced versus standard dose corticosteroids for patients with vasculitis. The authors reported, for the outcome of mortality, a reduction in deaths of 21 per 1000 with a 95% CI that includes a 60 per 1000 reduction, but also a 36 per 1000 increase (Figure 8). If the authors used an MID of 1% they would rate their certainty in an important effect. Given that the CI crosses the MID threshold they would rate down for imprecision.

Moreover, the width of this CI would prompt the review team to consider rating down twice for imprecision. Indeed because the CI not only crosses the MID for benefit, but also includes an important harm, they would rate down twice for imprecision. Thus, even before considering any other reason for rating down, the authors have only low certainty evidence that the lower dose regimen results in an important reduction in mortality (the target of their certainty rating).

Figure 8. An example of rating down two levels for imprecision in a systematic review of different doses of corticosteroids on mortality in patients with vasculitis. Since the confidence interval includes both important benefit and important harm, the review authors should consider rating down two levels for imprecision.
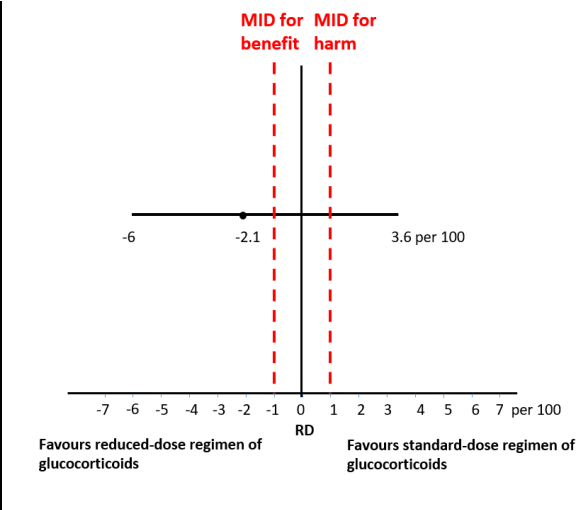
The second consideration that bears on the decision regarding rating down once or twice has to do with the most suitable plain language statement that accompanies the certainty of evidence. Consider the example of corticosteroids for patients with vasculitis that includes a CI ranging from a 6% reduction to a 3.6% increase. Assuming review authors do not have concerns on the other GRADE domains, would it convey the optimal message about certainty stating that the lower dose regimen likely results in an important reduction in mortality (the statement that would accompany rating down once) or that it may result in an important reduction in mortality (the statement that would accompany rating down twice). If the review authors considered the latter statement more appropriate (as in our view they should) they would rate down twice for imprecision. This highlights that it can be useful for Core GRADE users to consider what would be the most appropriate statement to communicate to their target audience.

The two considerations also apply to imprecision judgements when Core GRADE users choose the null as the threshold of interest. For example, consider a situation in which the Core GRADE users rate their certainty in a

benefit (threshold the null), and the CI also includes clearly important harm. The finding that the CI is consistent with both important benefit and important harm motivates a plain language summary stating that the intervention "may" result in a benefit and the rating down two levels for imprecision.

5. Rating down for imprecision when there is a large effect and limited sample size

When the CI crosses the threshold(s) of interest, Core GRADE users will rate down for imprecision and do not need to consider the sample size. However, if the CI does not cross the threshold and the effect is large they must consider that large effects are unusual in interventions tested in RCTs. Attempts to replicate results of early studies suggesting such effects often fail. Thus, we suggest that when the CI does not cross threshold(s) of interest and effects on binary outcomes are implausibly large (certainly relative risk reduction greater than 40%, possibly greater than 30%), Core GRADE users should consider rating down for imprecision if the sample size and number of events across all contributing studies are limited.

Our criteria for "limited" rely on routine sample size calculations that one would undertake when planning a single RCT (Appendix 5). For binary outcomes, these involve specifying the acceptable error rates: alpha (typically 0.05) and beta (typically 0.20), the control group event rate (chosen from the context), and a modest relative risk reduction (RRR), typically 20% or 25%. We call the sample size that emerges from the calculation the optimal information size (OIS). If the total sample size of all the studies included in a meta-analysis exceeds the OIS one does not rate down; if the total sample size proves less that the OIS one rates down for imprecision. There are many online calculators available to address the OIS (e.g. https://www.openepi.com/SampleSize/SSCohort.htm).

Core GRADE users can make the same calculation for continuous variables by specifying the smallest difference between intervention and control that one would want to avoid missing (i.e., the MID) if it indeed existed and using the standard deviation from one of the existing studies. An alternative, a rule of thumb, would suggest that to not have concerns about imprecision (i.e. to not rate down) would require 400 patients per group (total sample size 800). A prior GRADE article and Appendix 5 provide further details and examples of OIS exploration for both binary and continuous variables.[14]

6. Conclusion

The process of assessing the certainty of evidence requires choosing a threshold, either the null or the MID, and then choosing the target of certainty by noting the location of the point estimate in relation to the threshold. When the initial choice of threshold is the null, if the point estimate falls very close to this threshold, Core GRADE users rate certainty in little to no effect. For judging imprecision, if the confidence interval does not cross the threshold, Core GRADE users typically do not rate down for imprecision; if it crosses, they do. Core GRADE users may rate down twice when the confidence interval crosses more than one threshold, in particular when it crosses thresholds of important benefit and important harm. Finally, when the confidence interval does not cross the threshold, but the effect is large, they invoke the OIS and rate down for imprecision if the total sample size fails to meet the OIS criterion.

References

1. Gyuatt G, Agoritsas T, Brignardello-Petersen R, et al. Core GRADE 1: Overview of the Core GRADE Process *BMJ (in submission)* 2024

2. 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. doi: 10.1016/j.jclinepi.2017.05.006 [published Online First: 20170518]

3. Zeng L, Brignardello-Petersen R, Hultcrantz M, et al. GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings. *J Clin Epidemiol* 2021;137:163-75. doi: 10.1016/j.jclinepi.2021.03.026 [published Online First: 20210420]

4. Zeng L, Brignardello-Petersen R, Hultcrantz M, et al. GRADE Guidance 34: update on rating imprecision using a minimally contextualized approach. *J Clin Epidemiol* 2022;150:216-24. doi: 10.1016/j.jclinepi.2022.07.014 [published Online First: 20220804]

5. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10(4):407-15. doi: 10.1016/0197-2456(89)90005-6

6. Schünemann HJ, Neumann I, Hultcrantz M, et al. GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions. *J Clin Epidemiol* 2022;150:225-42. doi: 10.1016/j.jclinepi.2022.07.015 [published Online First: 20220805]

7. Guyatt G, Yao L, Murad MH, et al. Core GRADE 6: Presenting the evidence in summaries of findings tables. *BMJ (in submission)* 2024

8. Carrasco-Labra A, Devji T, Qasim A, et al. Minimal important difference estimates for patient-reported outcomes: A systematic survey. *J Clin Epidemiol* 2021;133:61-71. doi: 10.1016/j.jclinepi.2020.11.024 [published Online First: 20201213]

9. Zeng L, Yao L, Wang Y, et al. Presentation approaches for enhancing interpretability of patient-reported outcomes in meta-analyses: a systematic survey of Cochrane reviews. *J Clin Epidemiol* 2023;158:119-26. doi: 10.1016/j.jclinepi.2023.03.027 [published Online First: 20230405]

10. Zeng L, Li SA, Yang M, et al. Qualitative study of guideline panelists: innovative surveys provided valuable insights regarding patient values and preferences. *J Clin Epidemiol* 2023;161:173-80. doi: 10.1016/j.jclinepi.2023.07.014 [published Online First: 20230729]

11. Zeng L, Hultcrantz M, Tovey D, et al. Rating certainty when the target threshold is the null and the point estimate is close to the null. *BMJ Evidence-based Medicine (Accepted)* 2024

12. Stern A, Skalsky K, Avni T, et al. Corticosteroids for pneumonia. *Cochrane Database Syst Rev* 2017;12(12):Cd007720. doi: 10.1002/14651858.CD007720.pub3 [published Online First: 20171213]

13. Santesso N, Glenton C, Dahm P, et al. GRADE guidelines 26: informative statements to communicate the findings of systematic reviews of interventions. *J Clin Epidemiol* 2020;119:126-35. doi: 10.1016/j.jclinepi.2019.10.014 [published Online First: 20191109]

14. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. *J Clin Epidemiol* 2011;64(12):1283-93. doi: 10.1016/j.jclinepi.2011.01.012 [published Online First: 20110811]