GRADE SERIES
GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions

Holger J. Schünemann\textsuperscript{a,b,c,d,*}, Ignacio Neumann\textsuperscript{b}, Monica Hultcrantz\textsuperscript{e}, Romina Brignardello-Petersen\textsuperscript{b}, Linan Zeng\textsuperscript{b,f}, M Hassan Murad\textsuperscript{g}, Ariel Izcovich\textsuperscript{h}, Gian Paolo Morgano\textsuperscript{b}, Tejan Baldeh\textsuperscript{b}, Nancy Santesso\textsuperscript{a,b}, Carlos Garcia Cuello\textsuperscript{a,b}, Lawrence Mbuagbaw\textsuperscript{a,b}, Gordon Guyatt\textsuperscript{b,c}, Wojtek Wiernio\textsuperscript{c}, Thomas Piggott\textsuperscript{a,b}, Hans De Beer\textsuperscript{i}, Marco Vinceti\textsuperscript{j}, Alexander G. Mathioudakis\textsuperscript{k}, Martin G. Mayer\textsuperscript{l,m,n}, Reem Mustafa\textsuperscript{o}, Tommaso Filippini\textsuperscript{j}, Alfonso Iorio\textsuperscript{b,c}, Robby Nieuwlaat\textsuperscript{a,b}, Maura Marcucci\textsuperscript{b,c}, Pablo Alonso Coello\textsuperscript{p}, Stefanos Bonovas\textsuperscript{q,r}, Daniele Piovani\textsuperscript{t}, George Tomlinson\textsuperscript{s,t}, Elie A. Akl\textsuperscript{b,u}, for the GRADE Working Group

\textsuperscript{a}World Health Organization Collaborating Center for Infectious Diseases, Research Methods and Recommendations, Michael G. DeGroote Cochrane Canada & McMaster GRADE Centres; McMaster University, Hamilton, Ontario, Canada
\textsuperscript{b}Department of Health Research Methods, Evidence and Impact, McMaster University, 1280 Main Street West, Hamilton, L8S 4L8, Ontario, Canada
\textsuperscript{c}Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, L8S 4L8, Ontario, Canada
\textsuperscript{d}Escuela de Medicina, Facultad de Medicina y Ciencia, Universidad San Sebastián, Sede, Santiago, Santiago, Chile
\textsuperscript{e}Swedish Agency for Health Technology Assessment and Assessment of Social services (SBU), S:t Eriksgatan 117, Stockholm, 102 33, Sweden
\textsuperscript{f}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
\textsuperscript{g}Evidence-based Practice Center, Mayo Clinic, 200 1st Street, SW, Rochester, MN 55905, USA
\textsuperscript{h}Department of Internal Medicine, German Hospital, Buenos Aires, Argentina
\textsuperscript{i}Guide2Guidance, Utrecht, The Netherlands
\textsuperscript{j}CREAGEN—Environmental, Genetic and Nutritional Epidemiology Research Center, Section of Public Health, Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Modena, Italy
\textsuperscript{k}Division of Immunology, Immunity to Infection and Respiratory Medicine, School of Biological Sciences, The University of Manchester, & North West Lung Centre, Wythenshawe Hospital, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK
\textsuperscript{l}EBSCO Clinical Decisions, EBSCO, 10 Estes Silpwich, MA 01938, USA
\textsuperscript{m}Triad Hospitals, Cone Health, 1200 North Elm St, Greensboro, NC 27401, USA
\textsuperscript{n}Open Door Clinic, Cone Health, 319 N Graham Hopedale Rd, Burlington, NC 27217, USA
\textsuperscript{o}Division of Nephrology and Hypertension, Department of Internal Medicine, University of Kansas Medical Centre, 3901 Rainbow Blvd, MS3002, Kansas City, KS 61160, USA
\textsuperscript{p}Centro GRADE Barcelona, Instituto de Investigacion Biomedica (IIB Sant Pau), Barcelona, Spain
\textsuperscript{q}Department of Biomedical Sciences, Humanitas University, 20072 Pieve Emanuele, Milan, Italy
\textsuperscript{r}IRCCS Humanitas Research Hospital, 20089 Rozzano, Milan, Italy
\textsuperscript{s}Institute of Health Policy, Management, and Evaluation, University of Toronto, Toronto, Ontario, Canada
\textsuperscript{t}Biostatistics Research Unit, University Health Network, Toronto, Ontario, Canada
\textsuperscript{u}Department of Internal Medicine, American University of Beirut, P.O.Box 11-0236, Beirut, Lebanon

Conflict of interest: The authors declare no direct financial conflict of interest. They are members or contributors to the GRADE Working Group.

Author Contributions: Holger J Schünemann: Conceptualization and design; Writing—original draft, finalizing article; Ignacio Neumann: Conceptualization and design; Writing—Input on draft, final approval of article; Monica Hultcrantz, Linan Zeng, M Hassan Murad, Ariel Izcovich, Gian Paolo Morgano, Nancy Santesso, Tejan Baldeh, Lawrence Mbuagbaw, Carlos Garcia Cuello, Gordon Guyatt, Wojtek Wiernio, Thomas Piggott, Hans De Beer, Marco Vinceti, Alexander G. Mathioudakis, Martin G. Mayer, Reem Mustafa, Tommaso Filippini, Alfonso Iorio, Robby Nieuwlaat, Maura Marcucci, Pablo Alonso Coello, Elie A. Akl: Writing—critical review and input on draft, final approval of article; Romina Brignardello-Petersen, Stefanos Bonovas, Daniele Piovani, George Tomlinson: critical review and input on draft, statistical methods, final approval of article.

* Correspondence to: Michael G. DeGroote Cochrane Canada and McMaster Centres, Departments of Health Research Methods, Evidence and Impact of Medicine, McMaster University Health Sciences Centre, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada, Tel.: +1 905 525 9140 x 24931; fax: .

E-mail address: schuneh@mcmaster.ca (H.J. Schünemann).

https://doi.org/10.1016/j.jclinepi.2022.07.015
0895-4356/© 2022 Elsevier Inc. All rights reserved.
Abstract

**Objectives:** Grading of Recommendations Assessment, Development and Evaluation (GRADE) guidance to rate the certainty domain of imprecision is presently not fully operationalized for rating down by two levels and when different baseline risk or uncertainty in these risks are considered. In addition, there are scenarios in which lowering the certainty of evidence by three levels for imprecision is more appropriate than lowering it by two levels. In this article, we conceptualize and operationalize rating down for imprecision by one, two and three levels for imprecision using the contextualized GRADE approaches and making decisions.

**Methods:** Through iterative discussions and refinement in online meetings and through email communication, we developed draft guidance to rating the certainty of evidence down by up to three levels based on examples. The lead authors revised the approach according to the feedback and the comments received during these meetings and developed GRADE guidance for how to apply it. We presented a summary of the results to all attendees of the GRADE Working Group meeting for feedback in October 2021 (approximately 80 people) where the approach was formally approved.

**Results:** This guidance provides GRADE’s novel approach for the considerations about rating down for imprecision by one, two and three levels based on serious, very serious and extremely serious concerns. The approach includes identifying or defining thresholds for health outcomes that correspond to trivial or none, small, moderate or large effects and using them to rate imprecision. It facilitates the use of evidence to decision frameworks and also provides guidance for how to address imprecision about implausible large effects and trivial or no effects using the concept of the ‘review information size’ and for varying baseline risks. The approach is illustrated using practical examples, an online calculator and graphical displays and can be applied to dichotomous and continuous outcomes.

**Conclusion:** In this GRADE guidance article, we provide updated guidance for how to rate imprecision using the partially and fully contextualized GRADE approaches for making recommendations or decisions, considering alternate baseline risks and for both dichotomous and continuous outcomes. © 2022 Elsevier Inc. All rights reserved.

**Keywords:** GRADE; Guidelines; Imprecision; Statistical significance; Systematic reviews; Health technology assessment

1. Introduction

In the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach, imprecision, risk of bias, indirectness, inconsistency, and publication bias are the main domains that should guide the assessment of the certainty of the best available evidence in the context of systematic reviews, health technology assessments (HTA) or guidelines. A decade ago already, the GRADE Working Group provided initial detailed guidance for rating imprecision [1]. This guidance recommended that systematic reviewers judge precision as adequate if the 95% CI excludes a relative risk (RR) of 1.0 or if the 95% CI includes RR of 1.0 but the CI does not include appreciable benefit or harm and the total number of events or patients exceeds the optimal information size (OIS). In the case of guideline developers, they should judge precision as adequate if clinical or public health action would not differ if the upper and the lower boundary of the CI or certainty ranges represented the true effect; this requires setting thresholds (i.e., boundaries of effects that would influence a decision), and consideration of different degrees of context. GRADE guidance suggested rating down by one level in such circumstances.

Subsequently, the GRADE working group clarified that when rating certainty of the evidence, authors of systematic reviews, HTA or guidelines should be rating how certain they are that the true effect lies within a particular range or on one side of a threshold (Box 1) [2,9]. This emphasized the consideration of thresholds based on different degrees of context when rating the certainty of evidence, especially when assessing imprecision. The GRADE approach also allowed rating down by two levels if those judging the evidence had additional concerns about imprecision based on a very low number of events or participants across the studies constituting the body of evidence. However, GRADE guidance is not fully operationalized for rating down by two levels for imprecision. In our previous GRADE guidance, we suggested “When considering the certainty of evidence, the issue is whether the CI around the estimate of treatment effect is sufficiently narrow. If it is not, we rate down the evidence quality by one level (for instance, from high to moderate). If the CI is very wide, we might rate down by two levels.” We also stated “When there are very few events and CIs around both relative and absolute estimates of effect that include both appreciable benefit and appreciable harm, systematic reviewers and guideline developers should consider rating down the quality of evidence by two levels.” [1] In the prior guidance, our examples, such as “The point estimate of the risk ratio (0.96) suggests no difference, but the CI includes a reduction in likelihood of remission of almost half, or an increase in the likelihood of over 50% (95% CI: 0.56, 1.69).” described rating down by two levels based on relative effects when recommendations and decisions should be based on absolute effects. Thus, users of GRADE were not yet optimally supported in their decision to rate down by one or two levels. Box 1 describes prior guidance and other key concepts that are useful for understanding this article.

In addition, there are scenarios, in particular when only one or a few studies with few events and a small sample size across studies are available, in which lowering the certainty of evidence by three levels for imprecision is more
What is new?

Key Findings
- Grading of Recommendations Assessment, Development and Evaluation (GRADE) previously described rating imprecision based on the width of the confidence intervals, the optimal information size, and assumed range of effects that might influence a recommendation.
- It now provides guidance for how to rate imprecision using the partially and fully contextualized GRADE approaches for making recommendations or decisions, considering alternate baseline risks and for both dichotomous and continuous outcomes.

What this adds to what is known
- GRADE now provides additional guidance for rating imprecision using the partially and fully contextualized GRADE approaches for making recommendations or decisions, considering alternate baseline risks and for both dichotomous and continuous outcomes.

What is the implication, what should change now?
- Users of GRADE should use this approach for rating imprecision when using partially or fully contextualized approaches to rating certainty of evidence, that is, when evidence is used to make recommendations or decisions according to the corresponding GRADE guidance.

appropriate than lowering it by two levels. Consider the question if pneumatic compression devices or graduated compression stockings should be used in acutely and critically ill medical patients. The systematic review supporting the corresponding American Society of Hematology (ASH) guideline identified one randomized controlled trial in which one pulmonary embolism occurred in 43 participants [10]. The guideline panel questioned if it was sufficient to rate down by only two levels for imprecision considering both the relative and absolute effect. For pulmonary embolism the RR was 0.38 (0.02 to 8.86) with an absolute estimate of seven fewer per 1,000 (from 43 fewer to 342 more per 1,000) based on a baseline risk of 4.3% (43/1,000) from observational studies.

Although, this extremely imprecise evidence often results in a very low certainty rating for other reasons such as indirectness for the population (e.g., because rarely all populations are represented in a limited number of small studies or because the outcome is not measured at the appropriate length of follow-up). The uncertainty associated with this type of evidence may sometimes be mitigated by focusing on larger observational studies but with a trade-off for higher risk of bias. Regardless, the certainty rating should reflect the exact reasons (e.g., imprecision) why the certainty was lowered, as it may have important consequences for research (e.g., to increase the sample size and number of events in future studies if few events and few participants are the main concerns and rating down by three levels is justified). Until now, GRADE had not fully operationalized guidance to rate down for imprecision by two levels and rating down by three levels was not at all considered.

Furthermore, we did not specify when different or uncertain baseline risks may affect the rating of imprecision. Indeed, in simple models combing baseline risks and relative risks, absolute effect estimates will have narrower confidence intervals in a population with a very low baseline risk than a population with a high baseline risk, even if the relative risk and its confidence interval are the same.

Several recent developments paved the way for this new practical guidance for GRADE users on how to rate down for imprecision by one, two and three levels addressing the issues raised above: First, the work on terminology for GRADE’s rating down by three levels (initially only used for risk of bias in the context of using the ROBINS-I tool) second, the further operationalization of the partially contextualized GRADE approach for network meta-analysis [11,12]; third, the ongoing work on thresholds for the effect size in the GRADE Evidence to Decision Frameworks [13]; and fourth, prior work on defining the certainty of evidence and considering imprecision beyond the OIS acknowledging less emphasis on statistical significance [2,9,14]. GRADE defined the certainty of the evidence as the certainty that a true effect lies on one side of a specified threshold or within a chosen range of effect sizes to reflect GRADE’s recognition that evidence rating should support decision making [2]. We also acknowledged that the domains that determine the certainty have an impact on the shape and width of that certainty range and, thus, on decision-making [9]. In particular, the development of the GRADE Evidence to Decision Frameworks has had important impact on how one can conceptualize the thresholds that are necessary to make judgments if a confidence interval around an absolute intervention effect is too wide or sufficiently narrow [3]. These frameworks support structured decision-making and moving from evidence to a recommendation or decisions. Four criteria support balancing the health effects of interventions: 1) if desirable health effects (or benefits) are substantial; 2) if undesirable health effects (harms) are substantial; 3) the variability or uncertainty about values (i.e., the relative importance of the outcomes addressing the health effects); and 4) the certainty of the evidence in the health effects.
For the criteria relating to the desirable and undesirable health effects, decision makers, e.g., a guideline development group, judge if an intervention has a no or a trivial, a small, a moderate or a large effect (Box 1). Making these judgments by and across outcomes is important for understanding which interventions provide the greatest net health effect and together with the criterion on the variability and certainty about the relative importance of the outcomes (values) and the overall certainty of the evidence they determine the overall balance of the health effects [3].
2. Objectives

To operationalize and provide guidance for rating down for imprecision by one, two and three levels based on serious, very serious and extremely serious concerns, using the contextualized GRADE approaches for dichotomous and continuous outcomes to support decision-making, such as for guideline recommendations and HTA.

3. Methods

3.1. Overview

We identified, over several years working on systematic reviews and guidelines (e.g., ASH, World Allergy Organization, WHO), examples illustrating the limitations of: 1) the guidance to rate down for imprecision by no more than two levels; 2) the lack of operationalization of guidance for rating down by two levels; and 3) when baseline risks vary or are uncertain.

3.2. Examples

We asked the authors of this article to identify examples from systematic reviews and guidelines that could be used for discussion and to guide the approach to rating down by one and two levels and those in which the current GRADE guidance for rating imprecision seems to have inappropriately rated down by one or two levels only rather than three. The example about pneumatic compression devices vs. graduated compression stockings for acutely and critically ill medical patients [10] is listed in Appendix 1 together with another example.

3.3. Operationalization of guidance

We developed the approach through iterative discussions and refinement in online meetings and through email communication. The lead authors revised the approach according to the feedback and the comments received during these meetings and developed guidance for how to apply it. We presented a summary of the results to the relevant GRADE project group with an invitation to the GRADE Working Group as a whole in September 2021 (approximately 75 attendees) and all attendees of the GRADE Working Group meeting for feedback in October 2021 (approximately 80 attendees). The approach was formally approved by the GRADE Guidance Group on March 8, 2022.

4. Results

This guidance operationalizes rating down for imprecision by one, two and three levels based on serious, very serious and extremely serious concerns, using the partially and fully

---

**Box 2 Steps for rating imprecision using the partially contextualized approach**

Step 1. Define your outcome as dichotomous or continuous

Step 2. For the body of evidence, set thresholds for absolute effects of health outcomes that correspond to small, moderate or large effects, both desirable and undesirable (note, that outcomes, e.g., mortality, can be a desirable outcome if it is reduced or undesirable if it is increased): use existing evidence for these thresholds from other decision-makers, consensus by relevant stakeholders, empirical evidence about thresholds that integrate the absolute effect size and the relative importance of the outcomes or the best guess by content experts if nothing else is available. It is important to state how you set your thresholds.

Step 3. Choose the target of the rating of certainty of evidence in relation to those thresholds for the point estimate of the absolute effect. That is decide if you rate if the effect lies between two thresholds (i.e., between small desirable and undesirable effects or between small and moderate or moderate and large) or beyond the threshold for large effects or decide if you rate the certainty that the effect is beyond or below one of the thresholds.

Step 4. Calculate the absolute effect estimates for the body of evidence for the outcome of interest including its confidence intervals based on the baseline risk and relative effect or if relevant the meta-analytic estimate of the risk difference across studies (e.g., if there are very small number of events in the studies).

Step 5. Determine how many thresholds the confidence interval crosses regardless of if the effect estimates suggest a desirable or undesirable health effect (do not count the “no effect” as a threshold).

Step 6. Rate down by as many levels as thresholds are crossed.

Optional Step 7. If the effect is large (i.e., the point estimate falls beyond the threshold for large effects) and if it is based on an apparently small number of events or participants, consider using the review information size (RIS) by calculating the required sample size for small, moderate or large effects for the outcome of interest to determine if further rating down is required because the RIS is not met (see calculator in this article). Otherwise, use the rating in step 6. If the effect appears to be trivial or none (i.e., the point estimate falls within the threshold of trivial or no effect to small desirable and undesirable effects), check if the RIS for trivial to no effects is met, for example, for the assessment of equivalence of interventions, to determine if further rating down is required. Otherwise, use the rating in step 5.
contextualized GRADE approach. Many of the previous principles to rating the certainty of evidence for GRADE imprecision domain continue to apply [1]. However, the approach to rating the certainty of evidence for the partially and fully contextualized approaches has important new guiding principles that apply to both dichotomous and continuous outcomes. Generally, if the total number of events and participants is very small across the body of research evidence that provides estimates for a directly relevant outcome (e.g., with appropriate duration of follow-up and type of outcome), rating down for three levels appears more appropriate (see example from the ASH guidelines in the introduction and Figure 1 in Appendix).

The approach starts by identifying or defining thresholds for health outcomes that correspond to trivial or no, small, moderate or large effects as used in the GRADE EtDs for both the partially and fully contextualized approach. Rating imprecision in the partially and fully contextualized approach (e.g., a systematic review or HTA used in a guideline or coverage decision) should be based on absolute effects (which should be adjusted estimates in the case of non-randomized studies), and ratings of imprecision may differ for different baseline risk groups of the target population for an intervention or option. The thresholds are a result of combining the relative importance of the outcome (utility or value typically expressed on a scale of 0 to 1) and an absolute effect estimate (e.g., 10 fewer events/1,000 people of an undesirable outcome may be a small desirable effect or 10 more events/1,000 people of a desirable outcome also a small desirable effect, if the importance of the two outcomes is similar) [16,17]. Thus, two outcomes with similar importance should have similar thresholds, regardless of the type of question asked or health decision context. We begin by describing the rating of the certainty for the partially contextualized approach that includes the practical steps in Box 2.

4.1. Partially contextualized approach

The extent of rating down depends on the number of thresholds being crossed by the confidence interval. We also provide guidance for how to address imprecision about implausible large effects and trivial or no effects using the concept of the review information size (RIS) in the partially contextualized approach [9]. Given that the operationalization of the optimal information size (OIS) includes a consideration of an effect that is worthwhile focusing on one threshold, we use an alternative concept and term—the RIS—that is based on a similar mathematical approach but its definition does not rely on a worthwhile effect, which is counterintuitive to the purpose of using thresholds for different sizes of health effects [9].

4.2. Dichotomous outcomes

Raters should define the thresholds for categories of magnitude of effect. For desirable and undesirable health outcomes (also called benefits and harms and burden, respectively), using trivial or no effect, small, moderate, and large effects, results in a total of seven categories, with trivial or none being between the common thresholds of the small categories of desirable and undesirable health effects. Thresholds for desirable and undesirable health effects will usually be symmetric and typically a result of the absolute effect multiplied by the utility or the value assigned to the outcomes (or relative outcome importance) (Fig. 1). In general, thresholds for critical outcomes are expected to be smaller than those for important outcomes as defined by

---

**Fig. 1.** Describes the thresholds and ranges for trivial, small, moderate and large effects.
their value. The distance between the small and moderate thresholds and between the moderate and large thresholds may not be the same.

Although thresholds (the magnitude of effect that defines trivial or none, small, moderate or large effects) can be defined using various approaches, there are two key principles to follow: first, use absolute intervention effects to define the thresholds; and, second, thresholds should reflect the importance of the outcome. To achieve the latter, the process should involve key stakeholders (e.g., people affected by an intervention and providers) and consider available research (e.g., studies of values or utilities or qualitative research). For example, the threshold for a small beneficial effect (i.e., the magnitude at which a small effect would become trivial or no effect) should be smaller for mortality compared with headache. Systematic review authors lacking the relevant content knowledge can use GRADE’s imprecision guidance that is informed by

Fig. 2. Panel (A, B and C) describe a range of hypothetical scenarios to rating down for imprecision by one, two, and three levels respectively, while panel (D) describes the conceptual approach to not rating down. Please note that for demonstration purposes the confidence intervals are symmetrical while for most scenarios the confidence intervals will be asymmetrical. Conceptually, one should rate down by one level when the CI for the absolute effect crosses one of the thresholds across the range of categories of magnitude of effect (in our example, there are six thresholds for seven categories of magnitude of effect). One should rate down by two or three levels when the CI for the absolute effect crosses respectively two or three of the thresholds across the range of categories of magnitude of effect, respectively.

H.J. Schünemann et al. / Journal of Clinical Epidemiology (2022)
empirical thresholds for small, moderate and large effects; these empirical thresholds are emerging from research on how utilities and absolute effects lead to thresholds, require just an estimate of the utility to obtain the threshold and have been applied in guideline development [13,18]. They can also use their own thresholds based on informed guesses, those set by decision makers for specific outcomes in prior systematic reviews, HTA or guidelines, or determine them considering the number of events in relation to the number of participants in the studies included in the systematic review. It is important to be transparent about how the threshold was set. We will provide examples for the approaches.

4.2.1. Choosing targets of GRADE certainty of evidence ratings

After setting the thresholds, decision makers need to choose the target of the rating of certainty of evidence in relation to those thresholds, as deemed most relevant in their specific context. For example, they can decide to rate the certainty that the effect lies between two thresholds (i.e., between small desirable and undesirable effects or between small and moderate or moderate and large) or they can rate the certainty that the true effect lies beyond the threshold for large effects based on the point estimate which represents the most likely effect. They can also rate the certainty that the effect is beyond or below one of the
trivial, small, or moderate thresholds, in which case one would focus on the confidence intervals on one side only and determine how many thresholds are crossed.

Figure 2 panel A, B and C describe hypothetical scenarios to rating down for imprecision by one, two, and three levels respectively, although Figure 2 panel D describes the conceptual approach to not rating down for imprecision. Conceptually, across the range of categories of magnitude of effect, one should rate down by one level when the CI for the absolute effect crosses one of the thresholds (in our example, there are six thresholds for seven categories of magnitude of effect). For example, for the question if low molecular weight heparin (LMWH) or unfractionated heparin (UFH) be used in acutely ill medical patients for venous thromboembolism (VTE) prophylaxis, a systematic review for the ASH guidelines found nine studies reporting on the outcome mortality in the population [10]. The relative risk for the outcome mortality was 0.99 (95% CI 0.82 to 1.19). The corresponding absolute risk difference was one fewer per 1,000 (95% CI from nine fewer to 10 more per 1,000) and it was based on a baseline risk of 5.0% in the control arm. The certainty of evidence was downgraded by one level for imprecision because the confidence interval around the absolute effects just included the threshold between a small benefit to trivial and no benefit (assumed to be 10 per 1,000).

Along the same lines, one should rate down by two or three levels when the CI for the absolute effect crosses respectively two or three of the thresholds. For example, an ASH guideline panel examined the question if in patients receiving maintenance vitamin K antagonist therapy for treatment of VTE a longer (e.g., 6 to 12 weeks) international normalized ratio (INR) recall interval for anticoagulation or a shorter (e.g., four-week) INR recall interval during periods of stable INR control be used [19]. The systematic review team identified two studies that reported on the outcome major
bleeding. The RR for major bleeding in the population was RR 1.05 (95% CI 0.30 to 3.65) and the absolute risk difference was one more per 1,000 (95% CI from 12 fewer to 45 more). The baseline risk was 1.7% as per the mean annual risk reported across 11 RCTs. The panel downgraded the certainty of the evidence by two levels for imprecision because the confidence intervals for the absolute effect included small beneficial and harmful effects (also assumed to be 10 per 1,000).

For its living guideline on anticoagulation in patients with COVID-19 related acute illness, the ASH guideline panel considered the question if direct oral anticoagulants...

---

**Fig. 4.** If empirical estimates for small but important, moderate and large effects for continuous outcomes are available they should be used to define the relevant thresholds. If such estimates are not available, probably the most straightforward approach is to use standardized mean difference (SMD). Raters can use the guiding principles of thresholds for small (SMD = 0.2), moderate (SMD = 0.5), and large effects (0.8). In this example, more than three thresholds are crossed and one would rate down by three levels (although more thresholds are crossed).

---

**Fig. 5.** Legend: A systematic reviews of randomized trials includes a total of 100 participants and demonstrates a large effect that is considered implausible, e.g., an absolute mortality reduction of 10% from 20% to 10%. The RIS derived sample sizes would be 10,044 for being above the large (9%), 1,116 for being above the moderate (5%), and 496 for being above a small effect (2%) threshold. Raters should then check if the sample size is larger than 10,044 (do not rate down), 10,044 to 1,116 (rate down one level), 1,116 to 496 (rate down two levels) or less than 496 (rate down three levels). For details about the calculation and demonstrations see Appendix 2. Red font indicates rating down. Green font indicates that rating down is not indicated. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)
(DOACs), LMWH, UFH, Fondaparinux, Argatroban, or Bivalirudin at intermediate-intensity or at prophylactic-intensity should be used. The systematic review identified two studies reporting on the critical outcome ‘all-cause mortality’ in the population [20]. The OR for all-cause mortality was 2.21 (95% CI 0.69 to 7.03) and the corresponding absolute risk difference of 90 more per 1,000 (95% CI from 26 fewer to 322 more). The absolute risk was based on a baseline risk of 9.1% from the pooled mean event rate among studies that provided the best estimates of risk [20]. The CI for the absolute effect included the possibility of small benefit (10/1,000) and large harm, thereby crossing more than three thresholds. The certainty of evidence was downgraded by three levels for imprecision accordingly.

Not rating down for imprecision in trivial or no effects is appropriate if thresholds for small effects are not crossed (top, Fig. 2, panel D). For the question if LMWH or UFH prophylaxis should be used for patients undergoing major general surgery, an ASH guideline panel identified 34 studies reporting on major bleeding as a harmful effect [21]. The RR for major bleeding was an RR 0.97 (95% CI 0.78 to 1.20). The absolute risk difference was 0 fewer per 1,000 (95% CI from three fewer to three more) based on the best estimate for baseline risk of 1.5% in the control arm (UFH). Although the confidence interval did cross the point of no harm or benefit, it excluded the thresholds for small harm or benefit (assumed to be at 10/1,000) and the guideline panel considered the effects to be within the limits of equivalence for this outcome. Similarly, The European Commission Initiative on Breast Cancer guideline development group considered the question “Should screening using digital breast tomosynthesis in addition to digital mammography or digital mammography alone be used in organized screening programmes for early detection of breast cancer in asymptomatic women?”. They found fifteen studies that reported on interval cancers in the population group. The relative risk for the outcome was 1.03 (95% CI 0.97 to 1.09) and the absolute risk difference was two more per 100 (95% CI from two fewer to six more) based on a baseline risk of 70.8% in the comparison group. The certainty of the evidence was not downgraded because the confidence intervals for the absolute estimate effects included only the range for trivial or no benefit and harms (the threshold for a small effect assumed to be larger than 6/1,000) [22].

If a trivial or no effect is the most likely outcome based on the point estimate, crossing of the confidence interval of one, two or three thresholds will lead to rating down by one, two or three levels, respectively. If the certainty is rated that the effect is beyond or below one particular threshold (e.g., bigger than a trivial desirable effect), one needs to use that threshold to rate imprecision and adjacent thresholds (e.g., thresholds on the side of undesirable effects) to determine when to rate down by 1, 2 or 3 levels.

Figure 3 panel A shows another practical example from the Chilean National living COVID-19 guidelines. A meta-analysis of 10 randomized trials (n = 6,700) showed the use of tocilizumab may be associated with a lower mortality (RR 0.84, 95% CI 0.75-0.94) [23]. The guideline panel determined the thresholds for small, moderate and large effects as 1% (10/1,000), 2.5% (25/1,000) and 5% (50/1,000), respectively. When one obtains the absolute effects (risk differences) based on the baseline risk for a low risk population (i.e., 5% baseline risk), the point estimate (8 fewer/1,000) lies in the trivial or no effect range. Despite the statistical significance (which GRADE deemphasizes as the only domain to rely on to judge the certainty), the confidence interval of the risk differences crosses one threshold, i.e., 10/1,000, for the lower baseline risk (mild disease population, i.e., 5% baseline risk, 8 fewer/1,000, 95% CI from 9 to 38 fewer/1,000). If there was no concern regarding other GRADE domains (risk of bias, indirectness, inconsistency or publication bias), this finding would translate into the statement that tocilizumab has probably no or a trivial effect on mortality in low risk populations, based on moderate certainty of evidence (and not high certainty of evidence, given the imprecision that indicates small effects cannot be excluded) [24]. For the moderate disease population at a higher baseline risk (i.e., 15% baseline risk, 24 fewer/1,000 95% CI 9 to 38 fewer/1,000) the CI crosses two thresholds, i.e., 10/1,000 and 25/1,000. If there were no concerns on other domains, it would lower the overall certainty from high to low certainty.

Another example is the use of budesonide for COVID-19 for the reduction of hospitalization in non-hospitalized patients, which shows how, using this guidance, the number of thresholds crossed determines the degree of rating down: not rating down, rating down by two or three levels (Figure 3, panel B). It not only demonstrates that imprecision should not be rated on the basis of statistical significance alone, it also further demonstrates that certainty ratings are influenced by the baseline risk estimates.

This guidance now provides the systematic approach to provide different ratings for imprecision for different baseline risks, when appropriate. Note also that the width of the confidence intervals of absolute effects could also be influenced by imprecision of baseline risk estimates that could be factored in by using the boundaries of the confidence intervals of the baseline risk or uncertainty in the other GRADE domains (future GRADE guidance will address this issue) [9,25].

4.3. Continuous outcomes

If empirical estimates for small, moderate and large effects for continuous outcomes are available they should be used to define the relevant thresholds. For example, for the Chronic Respiratory Questionnaire, these estimates are changes of approximately 0.5, 1.0 and 1.5 on a seven point
If the realistic absolute effect is small, one should compare the sample size to the corresponding RIS. If the actual sample size is one, two, or three RIS thresholds off, rate down by one, two or three levels, respectively. A systematic review of randomized trials includes a total of 100 participants and demonstrates a large effect that is considered implausible, e.g., an absolute mortality reduction of 10% from 20% to 10%. The RIS derived sample sizes would be 10,044 for being above the large (9%), 1,116 for being above the moderate (5%), and 496 for being above a small effect (2%) threshold in each group. Raters should then check if the sample size is larger than 10,044 (do not rate down), 10,044 to 1,116 (rate down two levels), 1,116 to 496 (rate down three levels) or less than 496 (rate down three levels). The choice of the effect beyond the large thresholds to calculate the required sample size requires adding an (arbitrary) effect which pragmatically can be the small effect. That is for sample size calculations an 11% risk difference can be chosen. For details about the calculation and demonstrations see Appendix 2.
from previous guidance that erroneously stated 200 per

4.4.1. Implausible large effects

The RIS can be calculated in relation to large, moderate and small effect thresholds, i.e., based on corresponding absolute risk reductions or increases [1,9,30]. If the observed absolute effect is large (i.e., the point estimate is beyond the threshold of a large effect) one might want to determine if the actual review information (sample) size is larger than the required RIS for the large, moderate or small thresholds. The purpose is to determine if the effect is implausibly large, e.g., based on randomly high effect sizes. If raters choose to evaluate if an observed effect is implausibly large, GRADE’s guidance now suggests that if the RIS is smaller than that calculated for the large, moderate or small effect thresholds, the corresponding rating of imprecision should be lowered by three, two, or one level, respectively (Fig. 5). For example, consider a systematic review of randomized trials that includes a total of 100 participants and demonstrates a large effect that is considered implausible (e.g., an absolute mortality reduction or risk difference of 10% from 20% in the comparison to 10% in the intervention group). The RIS derived sample sizes would be 10,044 for being with precision above the large effect threshold (9%), 1,116 for being above the moderate (5%), and 496 for being above a small effect (2%) threshold (see Appendix 2 for the calculations). Raters should then check if the sample size is indeed larger than 10,044 (do not rate down), 10,044 to 1,116 (rate down one level), 1,116 to 496 (rate down two levels) or less than 496 (rate down three levels) (Fig. 5). Using these estimates, if an effect is believe to be implausible large, reduces the possibility of drawing conclusions about large effects with unduly high certainty. Note that the sample size will depend on the control event rate which should be determined using other available guidance [31,32]. Additional details for how to calculate the RIS are described in Appendix 2. We also have prepared an online calculator and graphical overview that can be accessed here (using the instructions and example in Appendix 2a and b): https://www.gradepro.org/calc/reviewinformationsize.

Note that for plausible large effects (i.e., effects that are expected to be large because of existing evidence) reliance on the RIS may lead to rating down by three levels and this approach should be used with caution. For example, the relative and absolute effects of a new oral anticoagulant for the prevention of stroke in patients with atrial fibrillation at risk would be expected to be large and penalizing
a body of evidence that has legitimate large effects may be inappropriate. Under these circumstances, one may follow a modified approach to calculate the RIS (see Appendix 2, scenario 1, option 4) which considers true effects that are expected to be plausibly large and provides smaller RIS estimates, thus avoiding to over-penalize this type of evidence.

4.4.2. Precision of trivial or small effects

Although the RIS (or previously the optimal information size) is predominantly helpful when determining if imprecision is of concern for implausible large effects, the RIS may also be helpful as a guiding principle by authors of systematic reviews. To guide those raters, the RIS could be used to estimate imprecision for trivial or small effects by setting arbitrary RIS thresholds for small, moderate and large effects. The RIS thresholds may be derived from indirect evidence (from other outcomes) or pragmatically by multiplying the RIS for small effects by 2.5 and 4 to derive the threshold arbitrarily for moderate and large effects (where 2.5 is derived from dividing the moderate by the small effect size 0.5/0.2 and 4 is derived by dividing the large by the small effect size from 0.8/0.2 in Cohen’s effect sizes).

If one obtains an estimate for the absolute effect that lies in the trivial to no threshold and it is realistic that the absolute effect is indeed trivial, one should compare the actual sample size to the corresponding calculated RIS to understand with what level of certainty one can say that the trivial effect is in fact not due to random error. If the actual sample size is smaller than the RIS for the one, two, or three adjacent thresholds, one should rate down by one, two or three levels, respectively. For example, if 10,044 participants are required for demonstrating a small (2%), 1,116 for demonstrating a moderate (5%), and 496 for demonstrating a large effect (9%), raters should check if the review sample size is indeed larger than 10,044 (do not rate down), 10,044 to 1,116 (rate down one level), 1,116 to 496 (rate down two levels) or less than 496 (rate down three levels) (Fig. 6). The same can be achieved by visually inspecting if the trivial effect that is observed crosses the thresholds for small, moderate or large effects and rating down by one, two, or three levels, respectively. This approach helps establishing with confidence that two interventions are equivalent. Although provided here as guidance, raters should proceed with caution, however, to not over penalize a body of evidence, for example when effects may be expected to be large or trivial based on other indirect evidence (see above).

Rating certainty in systematic reviews, health technology assessments and guidelines using the fully contextualized approach.

Our guidance can be used in systematic reviews, HTA and for guideline recommendations. However, ratings may differ when going from a systematic review to a decision or recommendation because of the existence of or assumptions about thresholds, the interplay of many outcomes that need to be balanced and the contextual factors (i.e., the factors in the EtD decision criteria) that may be used to develop thresholds and rate the certainty. These considerations represent the fully contextualized approach. For example, having estimated if desirable effects are large and undesirable effects are small based on thresholds that have been set before a decision is taken, may allow decision makers to better balance health effects. Box 3 describes the steps involved in rating imprecision using the fully contextualized approach.

Thresholds serve for the partially contextualized approach. A fully contextualized approach will consider other EtD decision criteria (e.g., cost) and the balance of all desirable and undesirable effects based on all outcomes which can further influence the rating of the certainty for imprecision and the overall certainty [9]. Box 3 describes the approach that is illustrated in Figure 7. It describes the interrelationship of the impact of imprecision of several outcomes to each other and final rating of the certainty focusing on desirable and undesirable health effects [9]. Using the guidance in Box 3, and a much simplified approach to the use of utilities, for the example in Figure 7, the aggregate acceptable number of severe infections is 20 based on the upper limits of the confidence intervals for the outcomes stroke, deaths and VTE and considering their relative importance (i.e., it is not a simple sum of 12 + 4 + 8 but includes the relative disutility which is greatest for death and stroke and comparably smaller for VTE). For instance, the disutility of death is 1.0, of stroke 0.8 and VTE 0.5, respectively. That amounts for stroke to 0.8 × 12 as the lower limit of the CI = 9.6, for death to 1.0 × 4 = 4, and for VTE to 0.5 × 8 = 4 for total of 17.6 of a utility weighted decision threshold across outcomes. For severe infections, the disutility is 0.7 and the upper limit of the CI is 12 = 8.4 which does not cross the decision threshold for net desirable effects. This threshold is not crossed by the outcome serious infections and one would not lower the certainty of the evidence for the outcomes that were initially rated down for imprecision. Note that the disutilities are approximate and used to enhance understanding of this example only.

4.5. How to obtain the thresholds

In ongoing work we are deriving empiric thresholds based on explicit utilities and absolute effects that can be used as guiding principles [13]. Another approach is deriving thresholds a priori with a guideline panel [18,23] which ideally should include the importance of the outcome considered. Thresholds that are derived during the decision making process with guideline development groups should be viewed with caution and well-reasoned by a decision making body. In this case, panelists should be appropriately trained to understand the meaning and implications of these decisions. In particular,
they should realize that both setting very small thresholds and/or very closely spaced ones has the consequence of increasing the likelihood of downgrading by imprecision. These are additional reasons why we recommend that thresholds should be decided after having rated the importance of each outcome and before evaluating the evidence. In general, these thresholds may require updating if new evidence about them emerges which, in turn, may require updating imprecision ratings for outcomes.

4.6. Relation to inconsistency

In some scenarios, the rating of the certainty is influenced by both imprecision and inconsistency. A meta-analysis of highly heterogeneous studies using the random-effects model usually leads to a wider confidence interval than that of the fixed effect model. In this case, rating down for both domains may not be necessary because imprecision could be caused by inconsistency and one has to carefully consider whether to rate down for both imprecision and inconsistency. For example, a meta-analysis showed that in patients with localized renal tumors, partial nephrectomy was associated with lower cancer specific mortality compared to radical nephrectomy [33]. The effect varied across studies (I square = 63%). The random effect estimate (HR 0.79; 95%CI, 0.57–1.11) was judged to be imprecise compared to the fixed effect estimate which was precise (HR 0.71; 95%CI, 0.59–0.85). In this case one might consider rating down for inconsistency only and avoid rating down for imprecision. On the other hand when meta-analyzing studies with very wide confidence intervals, it is unlikely that heterogeneity will be statistically appreciated, even if the point estimates of the individual studies vary considerably and suggest inconsistency may be present.

5. Discussion

In this updated GRADE guidance, we have conceptualized and operationalized how to rate imprecision for the partially and fully contextualized approach that serves for decision-making. We described scenarios that suggest lowering of the certainty of evidence for the imprecision domain by three levels (i.e., from high to very low) is reasonable. The approach also supports the use of thresholds relating to the categorical description of absolute effects in the GRADE EtDs (no to trivial, small, moderate or large effects). When recommendations or decisions are

---

**Box 3 Steps for rating imprecision using the fully contextualized approach**

Using the fully contextualized approach requires an evaluation of all outcomes using the partially contextualized approach first. To complete the fully contextualized rating follow these steps:

Step 1. Identify the desirable effects and their smallest plausible absolute effect sizes based on the limits of the confidence intervals.

Step 2. Identify the undesirable effects and their largest plausible absolute effect sizes based on the limits of the confidence intervals.

Step 3. Aggregate the smallest plausible desirable effects and, based on that, determine the largest plausible undesirable effects that would be acceptable to recommend the intervention (note that for a fully contextualized approach this may include criteria such as cost). Consider this overall threshold to determine if imprecision ratings of individual outcomes will have to be altered based on the largest plausible undesirable effects. Note also that if there is several undesirable health effects this may have to be done for each of them separately or in an aggregated way.

Step 4. Determine if the confidence intervals of the undesirable health effects overlap with the threshold for the acceptable plausible undesirable effects. If yes, precision ratings for the fully contextualized approach remain unchanged. Guideline panels will typically make a conditional recommendation because the balance between benefits and harms may not be certain, that is there are no clear net desirable health effects.

Step 5. If the threshold is not crossed, uncertainty based on imprecision may not influence the decision and lowering the certainty for imprecision for the desirable and undesirable effects may not be necessary to make a recommendation or decision (this will be rare) because there are net desirable effects. If the certainty of evidence for the body of evidence overall is then moderate or high, guideline panels often will make a strong recommendation. If the certainty of evidence for the body of evidence overall is very low or low guideline panels will usually make a conditional recommendation even if there are net desirable effects.

Note: Although we have not identified an example, hypothetically, even if there is no downgrading for any of the outcomes for imprecision using the partially contextualized approach, the cumulative uncertainty of all desirable or all undesirable outcomes together can make cumulative effects of desirable or undesirable outcome imprecise. If the cumulative uncertainty after combining all desirable or undesirable effects is so large that confidence intervals would cross a threshold, rating down for imprecision of one or more main outcomes may be justified.
made, the threshold should be defined a priori or based on transparent a posteriori judgments, primarily in the context of formulating recommendations, using the EtD frameworks. We provide guidance for both dichotomous and continuous outcomes and for varying baseline risks.

5.1. Strengths and limitations

This updated guidance builds on a decade of field testing of the original guidance, which has validated that original whose panels may find unrealistic, but observations from the field suggest that this can be done easily [18,23] or, in the future, empirically. The guidance is straightforward once thresholds are defined and is similar for both dichotomous and continuous outcomes. For the latter, the standardized effect sizes provide practical and transparent guidance for any continuous outcome without having to determine the thresholds.

Limitations include the risk of misuse of this guidance leading to excessive downrating and over-penalizing a body of evidence due to imprecision. Also, ideally, thresholds should be defined first which some guideline panels may find unrealistic, but observations from the field suggest that this can be done easily [18,23] or, in the future, empirically. However, although some GRADE members believe the process of obtaining thresholds early in the process will increase the time needed to develop a guideline, others believe setting thresholds before results are reviewed reduces unwanted uncertainty and inconsistency between panel members that lead to excessive discussion and, thus, makes the process more efficient. Our ongoing study will derive empirical suggested thresholds for when the relative importance (or utility value) of an outcome is known or can be assumed. In this case, with available absolute effects and their confidence intervals will allow applying this guidance in a straightforward manner. Another limitation is that the guidance using Cohen’s effect size for continuous effect sizes is based on assumptions and modeling about small, moderate and large effects [28]. Lastly, the CI of the absolute effect is a function of the baseline risk (i.e., this CI will widen as the baseline risk increases) but also influenced by the certainty one can place in that baseline risk (e.g., because of population indirectness or imprecision or both) [25]. This can make absolute estimates in high risk populations less precise and more uncertain compared to low risk population, and should be taken into consideration when thresholds are determined.

5.2. Implications for practice and research

Guideline groups and other decision makers should set effect thresholds a priori which can be done through review of research evidence, previous guidelines in similar topics developed with GRADE, surveys or empirical work such as the thresholds derived through an empirical controlled experiment we are currently concluding [13]. The guidance for continuous outcomes is already informed by thresholds for standardized mean differences which can be calculated for any continuous outcome if direct estimates are not available. The approach is particularly useful when using GRADE EtDs or the INTEGRATE framework [3,4,34]. For minimally contextualized approach, refer to the guidance developed for rating down by three levels [15].

There is a need to produce similar guidance for diagnostic questions and other types of bodies of evidence separately. We will develop guidance for rating down for three levels for indirectness depending on the degree of extrapolation and degree of uncertainty about the baseline risk estimates (which in turn also influences imprecision and the certainty of evidence of the baseline risk as demonstrated in Fig. 3) [25].

6. Conclusions

In this GRADE guidance article, we describe updated guidance for how to rate imprecision using the GRADE approach allowing the use of the full spectrum of levels of certainty from very low to high, considering different absolute effects and baseline risks for both dichotomous and continuous outcomes. This guidance describes the partially contextualized approach, which is when the judgments about the magnitude of effects are made, for example to categorize health benefits or harms in trivial or no, small, moderate or large effects on the GRADE EtD. It can be used in systematic reviews to inform decisions such as in guidelines or coverage decisions following a health technology assessment. It also provides the basis for rating imprecision with the fully contextualized approach [2]. It replaces the original guidance on rating imprecision albeit several of the prior concepts continue to apply [1].

Acknowledgments

Thank you to Leonardo Roever for feedback on the article. AGM was supported by the National Institute for Health Research Manchester Biomedical Research Center (NIHR Manchester BRC).

Appendix A

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.07.015.
References


