Journal of Clinical Epidemiology 137 (2021) 163–175

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings

Linan Zeng[a,b,*], Romina Brignardello-Petersen[b], Monica Hultcrantz[c], Reed A.C. Siemieniuk[b], Nancy Santesso[b], Gregory Traversy[d], Ariel Izcovich[e], Behnam Sadeghirad[b,f], Paul E. Alexander[b], Tahira Devji[b], Bram Rochwerg[b,g], Mohammad H. Murad[h], Rebecca Morgan[b], Robin Christensen[i,j], Holger J. Schünemann[b,g], Gordon H. Guyatt[b,g]

[a] *Pharmacy Department/ Evidence-based Pharmacy Centre, West China Second University Hospital, Sichuan University and Key Laboratory of Birth Defects and Related Disease of Women and Children, Ministry of Education, No. 20, Section 3, South Renmin Road, Chengdu, Sichuan, China, 610041*

[b] *Department of Health Research Methods, Evidence and Impact, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada, L8S 4L8*

[c] *Swedish Agency for Health Technology Assessment and Assessment of Social Services (SBU), S:t Eriksgatan 117, SE-102 33, Stockholm, Sweden*

[d] *Public Health Agency of Canada. 785 Carling Avenue, Ottawa, Ontario, Canada, K1A 0K9*

[e] *Internal Medicine Service, German Hospital, Pueyrred_on 1640, Buenos Aires C1118AAT, Argentina*

[f] *Department of Anesthesia, McMaster University, 1200 Main Street West Hamilton, Ontario, Canada, L8N 3Z5*

[g] *Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada, L8S 4K1*

[h] *Mayo Clinic Evidence-based Practice Center, Mayo Clinic, 200 1st Street SW, Rochester, MN, USA 55905*

[i] *Musculoskeletal Statistics Unit, the Parker Institute, Bispebjerg and Frederiksberg Hospital, University of Copenhagen, Nordre Fasanvej 57, DK-2000 Copenhagen F, Denmark*

[j] *Research Unit of Rheumatology, Department of Clinical Research, University of Southern Denmark, Odense University Hospital, J.B. Winsløws Vej 4, 5000 Odense C, Denmark*

Accepted 16 March 2021; Available online 20 April 2021

Abstract

Objective: To provide practical principles and examples to help GRADE users make optimal choices regarding their ratings of certainty of evidence using a minimally or partially contextualized approach.

Study Design and Setting: Based on the GRADE clarification of certainty of evidence in 2017, a project group within the GRADE Working Group conducted iterative discussions and presentations at GRADE Working Group meetings to refine this construct and produce practical guidance.

Results: Systematic review and health technology assessment authors need to clarify what it is in which they are rating their certainty of evidence (i.e., the target of their certainty rating). The decision depends on the degree of contextualization (partially or minimally contextualized), thresholds (null, small, moderate or large effect threshold), and where the point estimate lies in relation to the chosen threshold(s). When the 95% confidence interval crosses multiple possible thresholds (i.e., including both large benefit and large harm), it is not worthwhile for authors to determine the target of certainty rating.

Conclusion: GRADE provides practical principles to help systematic review and health technology assessment authors specify the target of their certainty of evidence rating.

Keywords: GRADE; Target of certainty of evidence rating; Thresholds; Evidence-based medicine; Systematic review; Health technology assessment

Conflicts of interest: None.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. BS reports receiving funding from PIPRA AG (www.pipra.ch) to conduct a systematic review and individual patient data meta-analysis on predictors of post-operative delirium in elderly for 2020-2021. BS also reports funding from Mitacs Canada, accelerate internship in partnership with Nestlé Canada to support his graduate student stipend from 2016 to 2018. Mitacs is a national, not-for-profit organization that has designed and delivered research and training programs in Canada working with universities, companies, and both federal and provincial governments. BS also reports funding from the International Life Sciences Institute (ILSI) - North America to support his graduate work for his 2015 academic year. In 2016-2017, BS worked part-time for the Cornerstone Research Group (CRG), a contract research organization. The ILSI funding and being employed by CRG are outside the required 3-year period requested on ICJME form. TD has received a CIHR knowledge synthesis grant and project grant for work on MIDs for PROMs. RC reported the Parker Institute, Bispebjerg and Frederiksberg Hospital is supported by a core grant from the Oak Foundation (OCAY-18-774-OFIL).

* Corresponding author. Telephone number: (86)-028-85503205, Fax number: (86)-028-85503716.

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

https://doi.org/10.1016/j.jclinepi.2021.03.026

What is new?

Key findings

- Systematic review and health technology assessment authors need to determine the target when they make GRADE ratings of certainty of evidence.
- This decision depends on the degree of contextualization, the threshold(s) chosen, and the relative position of point estimate in relation to the chosen threshold(s).
- When 95% confidence interval crosses multiple possible thresholds (e.g., including both large benefit and large harm), it is not worthwhile to determine the target of the certainty rating.

What this adds to what is known

- Building on prior GRADE guidance, this article provides specific suggestions for deciding on the target of certainty of evidence ratings.
- We provide practical guidance on how to make optimal choices regarding rating certainty of evidence using a minimally or partially contextualized approach.

What is the implication, what should change now

- The article will help systematic review and health technology assessment authors be aware of the importance of determining the target of their rating of certainty of evidence when using GRADE.
- Whenever they rate the certainty of evidence, systematic review and health technology assessment authors should be explicit about the target of the rating.

1. Background

1.1. GRADE guidance thus far

In previous guidance for authors of systematic reviews, health technology assessments, and clinical practice guidelines, the GRADE working group has offered clarification regarding how to make ratings of certainty (quality) of bodies of evidence [1]. In particular, when considering a choice between candidate interventions, ratings of certainty of evidence reflect our confidence that the true effect of an outcome lies on one side of a threshold (e.g., on the left side of the small effect threshold in Fig. 1) or within a chosen range (e.g., within the range of small effect in Fig. 1). In this article, written for systematic review and health technology assessment authors, we further clarify previous guidance.

In dealing with how to address the threshold or range, GRADE notes the importance of deciding on the associated level of contextualization. The choice of level of contextualization depends on what audiences would find most useful. In particular, for systematic reviews and health technology assessments, if audiences' focus is on whether there is an effect or whether there is an important effect, authors would choose a minimally contextualized approach. If audiences' focus is on the magnitude of effect (i.e., a trivial, small, moderate or large), authors would use a partially contextualized approach. Appendix 1 presents further semantic and conceptual issues related to minimally contextualized and partially contextualized approaches.

The minimally contextualized approach specifies two possible thresholds: no difference between groups (e.g., risk ratio [RR] of 1.0, risk difference [RD] of 0) or an important effect (i.e., minimal important difference [MID], also called small effect threshold) as a threshold for rating the certainty. Using a partially contextualized approach, reviewers would rate their certainty in relation to a range (i.e., a range of trivial, small, moderate or large effect) that is bounded by two thresholds (Fig. 1) [1,2,3].

Null effect and small effect threshold are possible thresholds for a minimally contextualized approach, while small, moderate and large effect thresholds provide the boundaries of ranges (i.e., a range of trivial, small, moderate or large effect) for a partially contextualized approach. A small effect threshold is also called a minimally important difference (MID).

1.2. Where is practical guidance still needed?

GRADE users still often fail to make an explicit statement about what it is in which they are rating their certainty (i.e., the target of the rating of certainty of evidence). This failure can have important consequences, because the judgments that influence the rating might depend on the choice of target.

For instance, consider a situation in which the RD in mortality between intervention A and placebo is 2 fewer deaths per 100 patients, with a 95% confidence interval (CI) from 0.5 fewer to 4 fewer death per 100 patients (Fig. 2). Some may rate their certainty that intervention A reduces mortality when compared to placebo (i.e., the target of certainty rating) and thus require no rating down for imprecision. Others might rate their certainty that there is an important difference in mortality between intervention A and placebo (i.e., the target of certainty rating). If they set the small effect threshold at 1% reduction of mortality, they would rate down for imprecision. Others may believe that users of their systematic review would be optimally informed by seeing both ratings.

Some may rate their certainty that intervention A reduces mortality when compared to placebo and not rate down for imprecision; Others may rate their certainty that intervention A has an important reduction in mortality compared with placebo and – depending on their threshold from importance, 1%, for example – would rate down for imprecision.

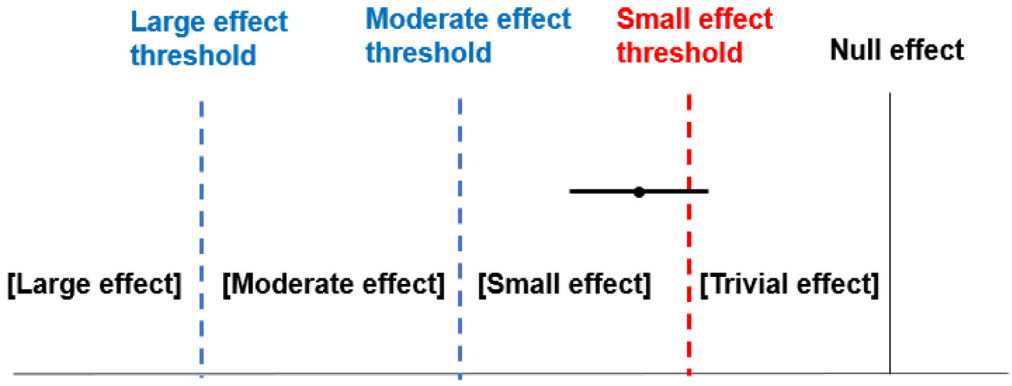


Fig. 1. Rating of certainty of evidence using a minimally or partially contextualized approach.

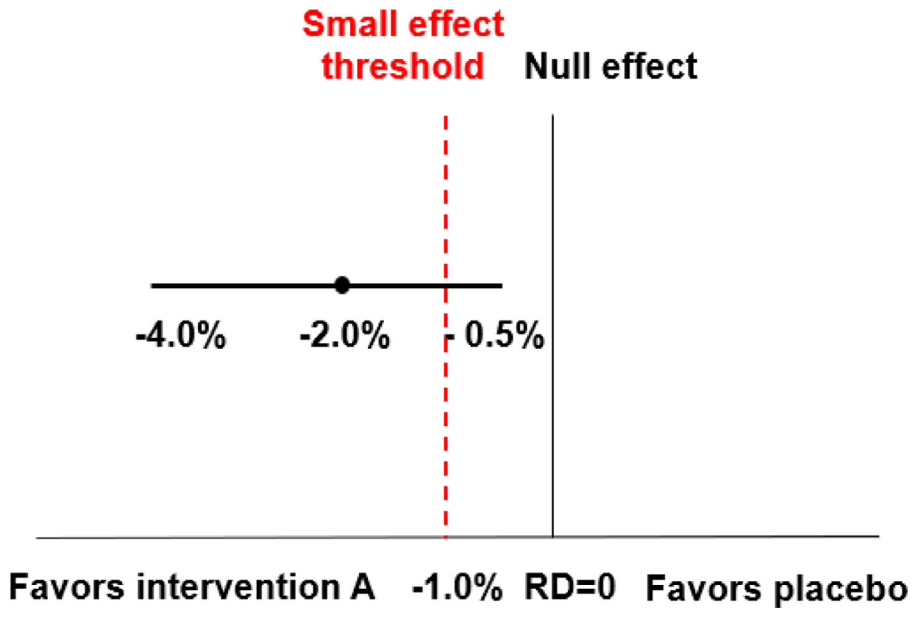


Fig. 2. A hypothetic example: intervention A versus placebo in reduction of mortality.

1.3. Scope of this article

This article provides practical principles and examples to help systematic review and health technology assessment authors make optimal choices regarding the target of their rating of certainty of evidence using a minimally or partially contextualized approach. In systematic reviews and health technology assessments, authors want to learn about the effect of interventions, and consider one outcome at one time. This article does not address issues that arise using a fully contextualized approach (e.g., a clinical practice guideline) or how the ratings would affect recommendations [4].

Further, being explicit about the basis for a threshold is important. However, the article will not address how to set thresholds other than the null. Choice of threshold will depend on the perspective (e.g., clinical perspective, public health perspective), the context (e.g., settings with more or less well-developed health care resources), and patients' values and preferences (e.g. the importance of the outcomes), among other factors.

2. **Practical principles for deciding about the target of rating of certainty of evidence**

We suggest four principles for choosing the target of rating the certainty of evidence and describe and illustrate these principles below. For simplification, we focus on single paired comparisons, and restrict the examples to situations in which results suggest a reduction in harmful outcomes (i.e., a reduction in the occurrence of the outcome is desirable). These principles, however, can be applied to ratings of certainty of evidence in all situations.

Principle 1 *Reviewers need to decide about the target of their certainty ratings.*

Given that one could rate one's certainty that the true effect lies on one or the other side of a threshold, or that it lies within a particular range (i.e., between two thresholds) (Fig. 1) [1], GRADE users should be explicit regarding the target of their certainty ratings.

In the hypothetic example of intervention A versus placebo on reduction of mortality (Fig. 2), when rating certainty in relation to the null effect, reviewers could specify they are rating their certainty that the true effect is a non-null effect (i.e., there is an effect), in this case a reduction in mortality greater than zero. Alternatively, when rating certainty in relation to the small effect threshold, they could specify they are rating their certainty that the true effect is important, in this case a reduction in mortality greater than the small effect threshold.

Principle 2 *The target of certainty ratings will depend on the degree of contextualization, the threshold chosen, and the point estimate.*

a *Degree of contextualization*

The decision regarding the target of certainty rating will differ depending on the degree of contextualization. For systematic reviews and health technology assessment authors, both minimally and partially contextualized approaches prove practical and useful.

a *The threshold choice*

When using a minimally contextualized approach, reviewers most often rate their certainty in relation to a single threshold. The null effect or a small effect threshold represent possible thresholds (Fig. 1) [1]. Rating certainty in relation to the former results in a rating of certainty in whether a non-null effect is present, while rating certainty in relation to the latter leads to a rating of certainty in whether an important effect is present.

When using a partially contextualized approach, one makes ratings in relation to a range. GRADE suggests four possible ranges (i.e., a range of trivial, small, moderate, or large effect) divided by three thresholds (i.e., small, moderate or large effect threshold) (Fig. 1) [1]. Choosing to rate certainty in relation to a particular range would result in a rating of certainty in whether a particular magnitude of effect is present.

When rating the certainty in relation to the null effect, one can present the threshold (i.e., the null effect) and effect estimates in either relative (e.g., RR = 1) or absolute terms (e.g., RD = 0). Depending on baseline risks, however, a particular relative effect may correspond with very different absolute effects. Thus, rating certainty in relation to threshold(s) other than the null effect requires presenting both threshold(s) and effect estimates in absolute terms [5]. Box 1 clarifies calculation of absolute risks from relative risks and baseline risks.

Systematic review and health technology assessment authors can enhance transparency by reporting, in preregistered protocols, the degree of contextualization and the particular threshold(s) or range(s) they will consider. For the threshold setting, as any threshold will involve some degree of uncertainty, authors could specify a range within which the threshold is likely to lie. Doing so, however, is likely to add considerable complexity to judgements, and is not something we would currently suggest.

Box 1 GRADE's approach of calculating an absolute risk from relative risk and baseline risk

We could calculate an absolute risk difference (RD) of an intervention versus a comparator using a risk ratio (RR) and an estimate of baseline risk (BR), with the following formula:

$$RD = BR\ (RR\text{-}1).$$

The baseline risk (BR) is the event rate in the comparator group (ranging from 0.00 to 1.00). We could obtain an estimated BR from population based observational studies or control groups of randomized controlled trials (RCTs) [6].

A RR below 1 represents a reduction in the risk of the event due to the intervention, and the RD could be presented as a negative number. A RR greater than 1 represents an increase in the risk of the event due to the intervention, and the RD would then be presented as positive. Alternatively, if one decided to frame the effect as an absolute risk reduction, the RD could be presented as positive.

We can also estimate the RD from an odds ratio (OR) with an estimate of BR, using the following formula [7]:

$$RD = [(1\text{-}OR)\ BR\ /\ \{1 - BR + (1\text{-}OR)\ BR\}]$$

c. *The point estimate*

For a given threshold, where the point estimate falls in relation to the chosen threshold(s) also determines the target of certainty rating. Figure 3 depicts the implications of particular point estimates when one chooses to rate certainty in relation to a small effect threshold. In situation (a), because the point estimate falls above the small effect threshold, one would rate one's certainty that the true effect is an important effect. In situation (b), however, as the point estimate falls below that threshold, one would rate one's certainty that the true effect is trivial or not important (i.e., smaller than the small effect threshold).

Still using this example (Fig. 3), one could choose a partially contextualized approach and rate certainty regarding whether a small effect is present (i.e., whether the true effect lies between small effect threshold and moderate effect threshold). In situation (a), because the point estimate falls within that range, one would rate certainty that the true effect is a small effect. In situation (b), as the point

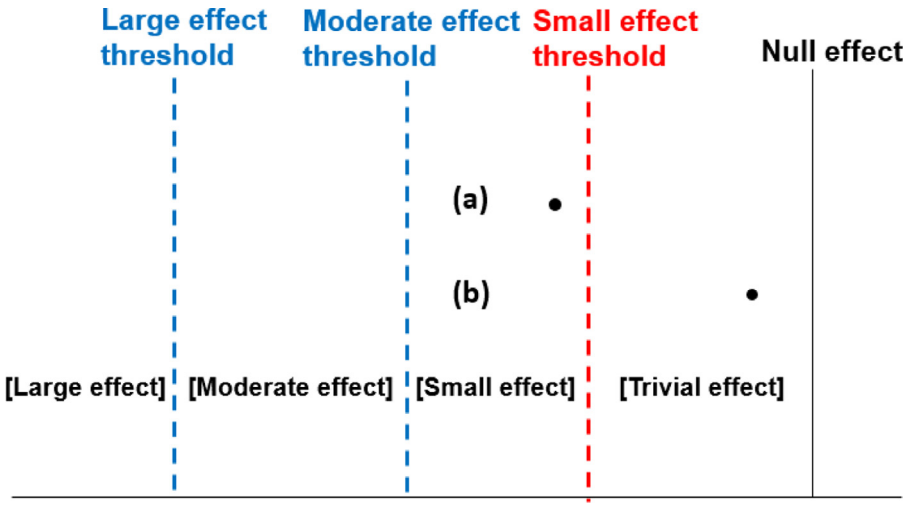


Fig. 3. The location of point estimate in relation to the chosen threshold(s) would determine the target of certainty rating (Principle 2c).

estimate falls below that range, one would rate certainty that the true effect is a trivial effect.

(a) Reviewers would rate certainty that the true effect falls above the small effect threshold (i.e., an important effect is present) or within the range of small effect (i.e., a small effect is present); (b) Reviewers would rate certainty that the true effect falls below the small effect threshold or below the range of small effect (i.e., a trivial effect is present).

When the point estimate is very close to the chosen threshold, one approach (Approach 1) would rate the certainty that the true effect is either above or below that threshold, depending on which side the point estimate falls. However close the point estimate is to the threshold, it will always be possible to carry the calculation to as many decimal places as necessary to determine on which side of the threshold the point estimate lies. Another approach (Approach 2) would rate certainty in relation to adjacent threshold(s). We clarify these two approaches using the following hypothetical example.

Consider a systematic review of intervention A versus placebo for prevention of stroke. Reviewers could had set a small effect threshold at 1 fewer stroke per 100 patients (Fig. 4a). The meta-analysis yields a RD of 0.99 fewer strokes per 100 patients, with a 95% CI from 0.1 fewer to 1.9 fewer strokes per 100 patients.

Approach 1: Although the point estimate is very close to the chosen threshold (i.e., the small effect threshold), reviewers could still rate certainty that the true effect lies below that threshold. In this case, the reviewer would rate their certainty that the effect is trivial (Fig. 4a).

Approach 2: Alternatively, reviewers could rate certainty in relation to two adjacent thresholds (i.e., the null effect, and a moderate effect threshold) (Fig. 4b). As the point estimate falls within the two thresholds, they would rate certainty that the true effect is a trivial or small effect.

Reviewers might be more comfortable with applying Approach 2 to situations in which the point estimate lies at, or is very close to, the null effect. It is not possible to rate certainty in point estimates alone (i.e., that there is no effect). Instead, one would rate certainty in relation to a range of trivial effect between a small effect threshold for benefit and a small effect threshold for harm (Fig. 4c). Because the point estimate falls within the range of trivial effect, reviewers would rate certainty that the true effect is a trivial to null effect.

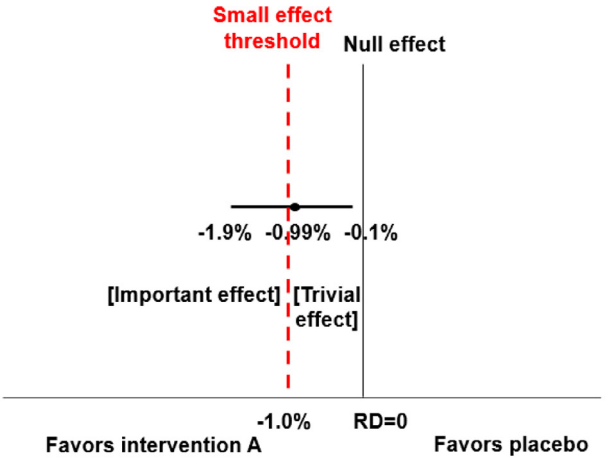
(a) Still rating certainty rating in relation to the chosen threshold (i.e., the true effect is trivial).
(b) Rate certainty in relation to two adjacent thresholds (i.e., the true effect is trivial or small).
(c) Rate certainty in relation to two adjacent thresholds (i.e., the true effect is trivial to null).

Principle 3 *Using a particular degree of contextualization, where the reviewers set the threshold(s) will determine the target of the certainty rating.*

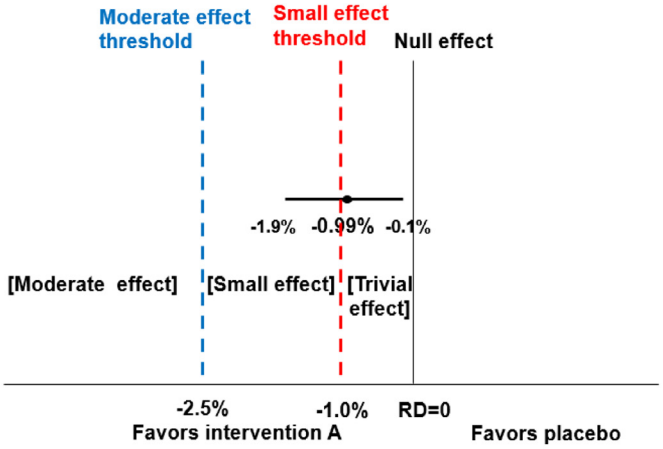
As presented in Figure 5, using a minimally contextualized approach and rating certainty in relation to a small effect threshold, if reviewers choose threshold 1, they would rate certainty that the true effect is larger than the small effect threshold (i.e., the true effect is an important effect). If they set the threshold at threshold 2, they would rate certainty that the true effect is smaller than the small effect threshold (i.e., the true effect is trivial).

When choosing threshold 1, reviewers would rate certainty that the true effect falls above that threshold; when choosing threshold 2, they would rate certainty that the true effect falls below that threshold.

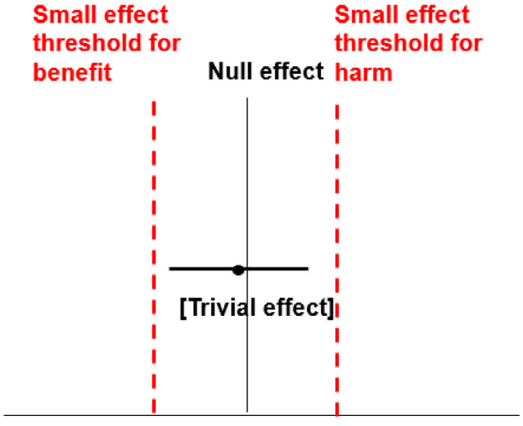
Principles 1 to 3 represent fundamental rules for deciding the target of certainty of evidence rating. There is, however, an exception. Consider Figure 6a. Here, considering only the point estimate, one could rate certainty that the true effect is greater than the small effect threshold. Because the 95% CI is so wide that we have very little



(a) Still rating certainty rating in relation to the chosen threshold (i.e. the true effect is trivial).



(b) Rate certainty in relation to two adjacent thresholds (i.e. the true effect is trivial or small).



(c) Rate certainty in relation to two adjacent thresholds (i.e. the true effect is trivial to null).

Fig. 4. Options for determining the target of the certainty rating when the point estimate is very close to the chosen threshold.

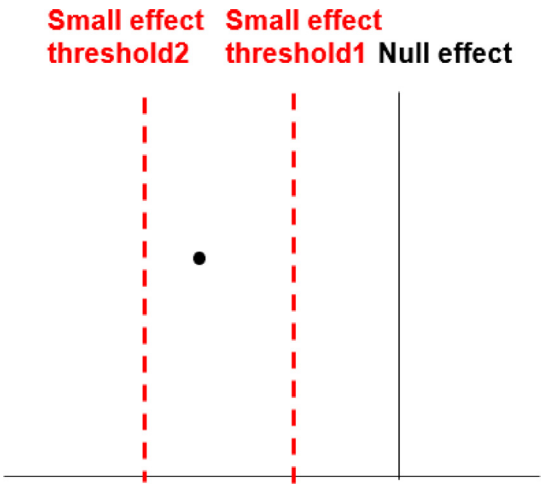


Fig. 5. Where reviewers set the threhold(s) will determine the target of the certainty rating (Principle 3).

idea of where the true effect lies, this would, however, make little sense. The true effect might represent a very large benefit, or a very large harm, thus, Principle 4:

Principle 4 *When the 95% CI crosses multiple possible thresholds, it is not worthwhile to choose a particular threshold and hence not worthwhile to decide about the target of the rating of certainty of evidence.*

Under these circumstances (Fig. 6a), rather than rating the certainty of evidence in relation to particular threshold(s), an appropriate conclusion would be reviewers have little idea of the true effect. Reviewers might also make this conclusion if the 95% CI included a large benefit and a small harm (Fig. 6b). In such situations, reviewers would rate down certainty of evidence by at least two levels. Exactly how wide the 95% CI has to be before reviewers abandon being explicit about the target of the certainty rating remains a matter of judgment.

In a partially contextualized approach one may (or may not) have specified boundaries of large benefit and large harm at the outset; in the minimally contextualized approach one would not. This does not preclude applying this principle in the minimally contextualized approach. Reviewers, using a minimally contextualized approach, can still make intuitive judgments regarding how wide the 95% CI has to be before abandon being explicit about the target of the certainty rating.

(a) The 95% CI crosses both threshold for large benefit and threshold for large harm; (b) The 95% CI crosses the threshold for large benefit and the threshold for small harm.

3. Application of principles

In this section, we describe how the principles presented above influence the judgments when rating the certainty of evidence. We focus on a single GRADE domain——imprecision, assuming no serious limitation in the other four domains (i.e., risk of bias, indirectness, inconsistency, and publication bias).

Appendix 2 presents further discussion regarding how one could consider the target of the rating of certainty of evidence, when simultaneously considering limitations in other GRADE domains [3,8].

Consider the hypothetical evidence in Figure 1.

(1) If reviewers are interested in the certainty of whether there is an effect, they would choose a minimally contextualized approach and rate certainty in relation to the null effect. As the point estimate falls above that threshold, they would rate certainty that there is an effect. Because the 95% CI does not cross the null effect, they would not rate down for imprecision.

Using the alternative within the minimally contextualized approach, reviewers would determine their cer-

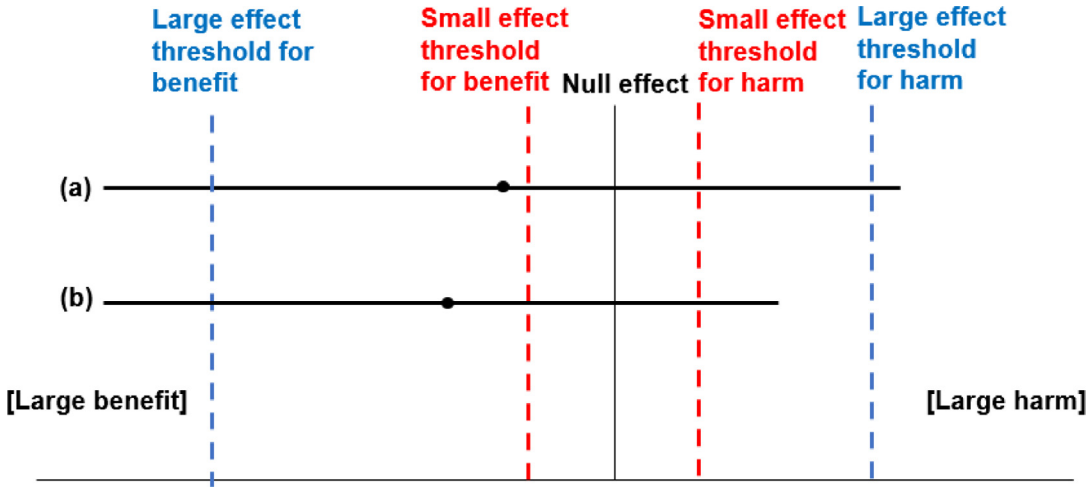


Fig. 6. When 95% CI is extremely wide, it is not worthwhile to decide about the target of the rating of certainty of evidence reviewers are very uncertain where the true effect lies (Principle 4).

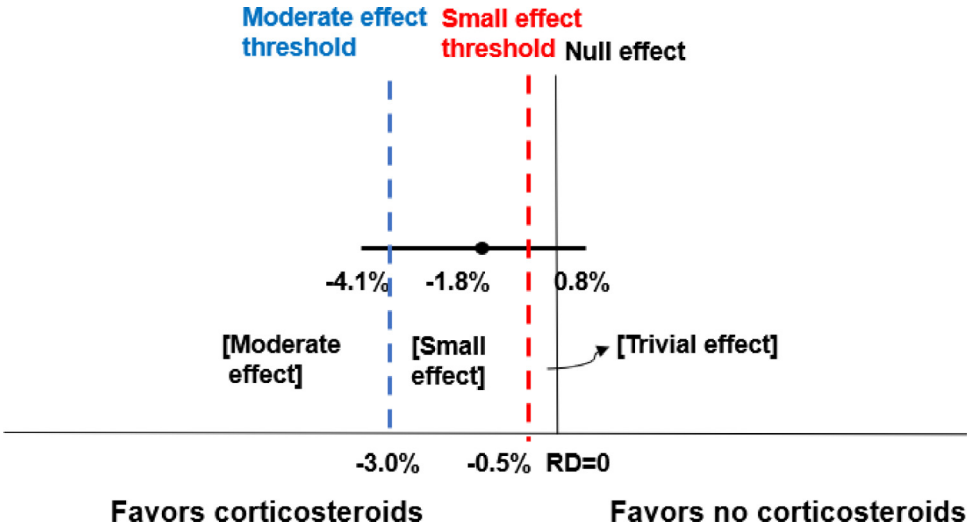


Fig 7. Application of principles in a systematic review of corticosteroids versus placebo in patients with sepsis.

tainty regarding whether an important effect exists, and would make their rating in relation to the small effect threshold. As the point estimate falls above that threshold, they would rate certainty that the effect is a small effect and would rate down for imprecision because the 95% CI crosses the chosen threshold.

(1) If interest lies in the certainty of whether a small effect exists, using a partially contextualized approach, reviewers would rate certainty in relation to the small effect range. In this case, because the point estimate falls within that range, the rating would be in relation to a small effect. As the 95% CI crosses the small effect threshold and it therefore remains plausible that the effect is trivial, one would rate down for imprecision.
(2) Using a partially contextualized approach, reviewers could rate the certainty that the effect is either small or trivial (i.e., it is smaller than moderate). If so, they would not rate down for imprecision because the 95% CI excludes values above the moderate effect threshold.

The chosen degree of contextualization and threshold(s) will depend on the audience. After review authors provide their rating, users can adjust ratings according to their own thresholds. For example, review authors could provide a rating that there is a any benefit and not rate down for imprecision if the 95% CI excludes the null effect. A user might then make one's own rating in whether the true effect is small and rate down for imprecision if the 95% CI crosses the small effect threshold, or that the true effect is smaller than a large effect and not rate down for imprecision if the 95% CI does not cross the large effect threshold.

4. Real examples

We present an example from a published systematic review [9] to illustrate the application of the principles. Authors of reviews did not always follow the guidance we suggest here – thus illustrating the desirability of the clarifications we present in this article. Appendix 3 presents more examples from published systematic reviews to illustrate the application of principles in this paper for continuous outcome, situation in which results suggest an increase in harmful outcomes and when the 95% CI is very wide.

Example 1: Determining the target of certainty rating using a minimally contextualized approach

Consider a systematic review of corticosteroids for patients with sepsis. A meta- analysis of 36 randomized controlled trials (RCTs) including 9,433 patients shows that corticosteroids yield 1.8 fewer deaths per 100 patients, with a 95% CI from 4.1 fewer to 0.8 more deaths per 100 patients (Fig. 7) [9]. Using a minimally contextualized approach and rating certainty in relation to the null effect, the authors of the review could have rated their certainty that corticosteroids reduced mortality (i.e., there is an effect). Because the 95% CI crosses the null effect, they would then have rated down the certainty due to imprecision [9].

Alternatively, still using a minimally contextualized approach, the authors could have rated their certainty in relation to a small effect threshold. For example, had they set the small effect threshold at 0.5 fewer death per 100 patients, they could have rated their certainty that corticosteroids result in an important reduction in mortality. If, however, they had chosen mortality reduction of 2 per 100 patients as a small effect threshold, they could have rated their certainty that corticosteroids have a trivial reduction in mortality. Wherever they set the threshold, they would have rated down for imprecision due to the overlap of the 95% CI with the small effect threshold.

Example 2: Determining the target of certainty rating using a partially contextualized approach

We continue with the systematic review of corticosteroids for treatment of patients with sepsis, this time using a partially contextualized approach. The authors might have set a small effect threshold at 0.5 fewer deaths per 100 patients, and a moderate effect threshold at 3 fewer deaths per 100 patients (Fig. 7). The authors could have then rated their certainty that corticosteroids result in a small reduction in mortality and rated down for imprecision because the 95% CI crosses both the small effect and moderate effect thresholds.

Using a minimally contextualized approach, the authors could rate their certainty that corticosteroids have an effect or have an important effect in reduction of mortality, and would rate down for imprecision in both ratings. Using a partially contextualized approach, they could rate their certainty that corticosteroids have a small reduction in mortality and would rate down for imprecision.

5. Conclusion

This article has, using hypothetical and real examples, built on prior GRADE guidance regarding rating certainty of evidence and provided specific suggestions for deciding on the target of certainty rating, and how to make judgments regarding rating down for imprecision [1]. The guidance is likely to be helpful to systematic review and health technology assessment authors who should take on the challenge of being explicit regarding thresholds or ranges that underlie their ratings of certainty of evidence.

Author contributions

Gordon H. Guyatt, Romina Brignardello-Petersen, Linan Zeng, Nancy Santesso, Monica Hultcrantz, and Reed A.C. Siemieniuk conceived and discussed the draft principles in this paper. All authors participated in the iterative discussion of the principles and examples in this manuscript. Linan Zeng wrote the first version of this paper. All authors reviewed and revised the paper.

Acknowledgment

The authors would like to thank GRADE Working Group members who gave valuable comments and suggestions on the draft of this article. The authors also thank attendants of GRADE Working Group meetings who have contributed to the article during group discussions.

Appendix 1. Minimally and partially contextualized approaches

1 Minimally contextualized approach

In this article, we combined what we previously called a non-contextualized approach in which reviewers use the null effect as a threshold to make judgments about the certainty of the evidence [1] with a partially contextualized approach in which reviewers use a small effect threshold, also called minimally important difference, as a threshold. We now use the label minimally contextualized approach to refer to both these approaches in which reviewers could use either the null effect or a small effect threshold.

The reasons for moving from the label of non-contextualized approach (in which no value judgments are needed) to the label of minimally contextualized approach are as follows:

First, a value judgment takes place when choosing the outcomes included in the systematic review or health technology assessment. By selecting important outcomes, those rating the certainty of evidence inevitably make a value judgment.

Second, even when rating the certainty of evidence in relation to the null effect threshold reviewers may need to make a value judgement. For example, when the point estimate is very close to the null effect, it is impossible to rate certainty of evidence that there is a trivial to null effect without setting a threshold different than the null. To rate certainty that the true effect is trivial, reviewers need to make a value judgment and set a small effect threshold for benefit and a small effect threshold for harm. These two thresholds form a range of trivial effect (See Fig. 4c in the main text).

Considering the points raised above, there is little rationale for the label non-contextualized in which no value judgment is needed.

2. Partially contextualized approach

Using a partially contextualized approach, reviewers could rate certainty that the true effect for a particular outcome, expressed in absolute terms, lies within or without the range of trivial, small, moderate or large effect [1].

In contrast to minimally contextualized approach, we are using the same conceptualization for partially contextualized approach in this article as in the previous GRADE guidance [1].

Reference

1. Hultcrantz M, Rind D, Akl EA et al. The GRADE Working Group clarifies the construct of certainty of evidence. Journal of clinical epidemiology 2017;87: 4-13.

Appendix 2. How reviewers would apply the principles to decide about the target of certainty of evidence rating when considering uncertainty from the five GRADE domains of limitations

1 Concept of certainty range

The GRADE approach for rating certainty of evidence includes five domains for rating down certainty (i.e., risk of bias, inconsistency, indirectness, imprecision, and publication bias). GRADE uses the concept of certainty range to characterize uncertainty that considers all the five domains of limitations [1,2]. The uncertainty associated with one of these domains of limitations, imprecision, can be quan-

tified by examining confidence intervals (CI) (or credible intervals for Bayesian analysis) [1,2]. The extent of uncertainty associated with the other four domains of limitations is not, thus far, amenable to quantification. Therefore, we still do not know how to quantify the certainty range.

2. **Application of the principles when considering the certainty range**

Conceptually, the other four GRADE domains extend and modify the distribution function of uncertainty for the best estimate of effect beyond that defined by the 95% CI [2]. The width of the certainty range would depend on the extent of concerns regarding the other four domains: the greater the concerns, the less is known about the width and shape of the probability distribution of the estimates in that range.

For example, consider a situation in which reviewers have serious concerns regarding risk of bias and indirectness, but have no reason to believe that the 95% CI would widen more on one side than the other (in other words, both risk of bias and indirectness can be acting in both directions). As presented in Appendix Figure 2.1a, the point estimate would not change, and the certainty range would widen beyond the 95% CI due to risk of bias and indirectness. Using a minimally contextualized approach, because the best estimate still suggests an effect greater than the small effect threshold, reviewers would rate certainty that the true effect is small, and would rate down certainty for risk of bias and indirectness but not for imprecision.

In another situation, if the risk of bias and the indirectness have a clear direction, the certainty range would widen in only one direction and the best estimate which presents the most probable true effect might move from one side of the threshold to another. As presented in Appendix Figure 2.1b, if reviewers are confident that risk of bias and indirectness overestimated the treatment effect and the best estimate should be moved from the left side of the threshold to the right side, they would rate certainty that the true effect is a trivial effect, and would rate down their certainty of evidence for risk of bias and indirectness but not for imprecision. Situations in which reviewers are aware of the direction of bias, and have a clear enough sense of its magnitude to confidently move the point estimate, are currently few and far between.

Conceptually, if reviewers could quantify the certainty range, when the uncertainty associated with some or all of the five domains of limitation is great that the certainty range becomes extremely wide, reviewers could abandon being explicit about the target of certainty rating (Appendix Fig. 2.1c).

(a) Reviewers are unaware of the direction to which the concerns of the other four domains of limitation would widen the certainty range than the 95% CI

(b) Reviewers are aware of the direction to which the concerns of the other four domains of limitation would widen the certainty range than the 95% CI

(c) The certainty range is extremely wide, reviewers could abandon being explicit about the target of certainty rating

References

1. Tikkinen KAO, Craigie S, Schünemann HJ, *et al.* Certainty ranges facilitated explicit and transparent judgments regarding evidence credibility. Journal of Clinical Epidemiology 2018;104: 46-51.

2. Schünemann JH. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? Journal of Clinical Epidemiology 2016;75:6-15.

Appendix 3. Application of principles in real examples

Here we present additional examples from published systematic reviews to illustrate how reviewers could use the principles in determining the target of certainty of evidence rating. The authors of these reviews did not always follow the guidance we suggest here – thus illustrating the desirability of the clarifications we present in this article.

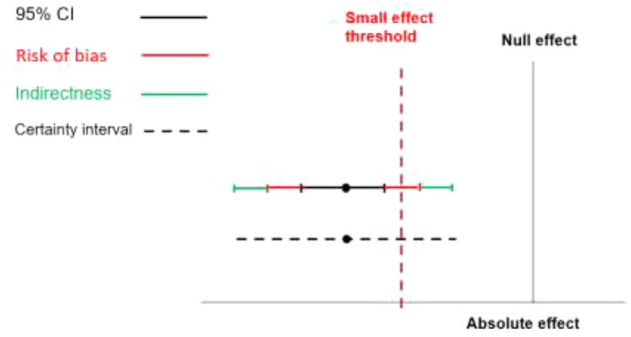
Example 1: Determining the target of certainty of evidence rating for continuous outcome

A systematic review of knee arthroscopy versus conservative management including 1,231 patients with degenerative knee disease from 10 randomized controlled trials (RCTs) yields a pooled mean difference in change of pain score from baseline until 3 month of 5.4 points higher on a 100-point scale (the higher score the better), with a 95% confidence interval (CI) from 1.9 to 8.8 points higher (Appendix 3 Fig. 3.1) [1]. The authors could have used a minimally contextualized approach and rated their certainty in relation to the null effect, in which case there would have been no need to rate down for imprecision.

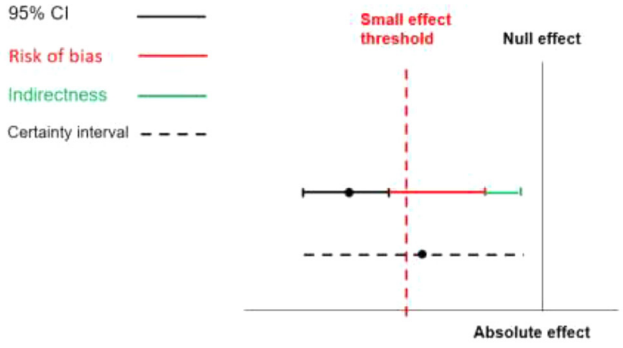
The authors, however, chose to rate certainty in relation to what they defined - based on a systematic review of minimally important difference (MIDs) offered for the relevant pain instrument - as an MID of 12 points (i.e., small effect threshold) [2]. Considering the point estimate, the authors rated their certainty that the true effect is smaller than the MID (i.e., the true effect is trivial). Further, because the 95% CI did not cross the threshold of 12, they did not rate down the certainty for imprecision. The systematic review authors found no other limitations in the body of evidence, and thus concluded with high certainty that knee arthroscopy does not result in an important reduction in pain when compared with conservative management in patients with degenerative knee disease [1].

Example 2: Determining the target of certainty of evidence rating when effect estimates suggest an increase in harmful outcome

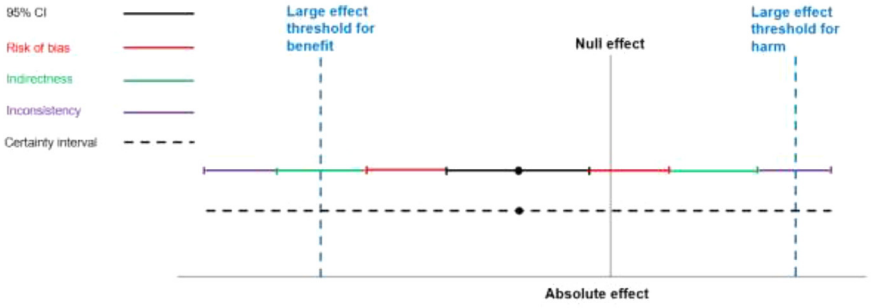
Consider a systematic review of clopidogrel and aspirin (dual antiplatelet therapy) versus aspirin alone in patients with minor ischaemic strokes or at high-risk of transient



(a) Reviewers are unaware of the direction to which the concerns of the other four domains of limitation would widen the certainty range than the 95% CI



(b) Reviewers are aware of the direction to which the concerns of the other four domains of limitation would widen the certainty range than the 95% CI

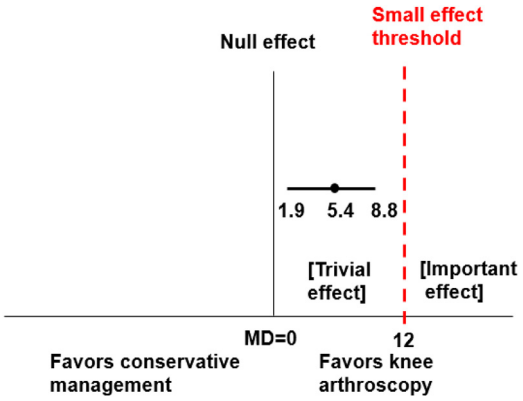


(c) The certainty range is extremely wide, reviewers could abandon being explicit about the target of certainty rating

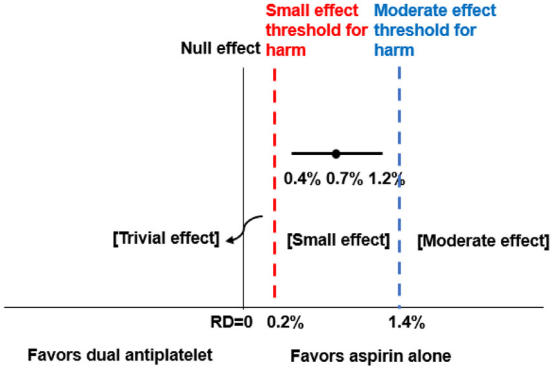
Appendix 2.1. Figure 1 Determine the target of certainty rating when consider the certainty range.

ischaemic attacks [2]. The point estimate of the risk difference (RD) for extracranial bleeding from the pooled analysis of 10,075 patients in three studies is 0.7 more bleedings per 100 patients with a 95% CI from 0.4 more to 1.2 more bleedings per 100 patients in dual antiplatelet therapy compared with clopidogrel alone (Appendix 3 Fig. 3.2) [2].

Using a partially contextualized approach, had the authors set the small effect threshold for harm at 0.2 more bleedings per 100 patients and the moderate effect threshold for harm at 1.4 more bleedings per 100 patients, the authors could then have rated their certainty that dual antiplatelet therapy results in a small increase in extracranial bleeding, compared with aspirin alone. Because the 95% CI does not overlap either with the small or the moderate harm thresholds, they would not rate down the certainty due to imprecision. As there were no other limitations of



Appendix 3.1. Figure 1 Determining the target of certainty rating for a continuous outcome (Example 1).

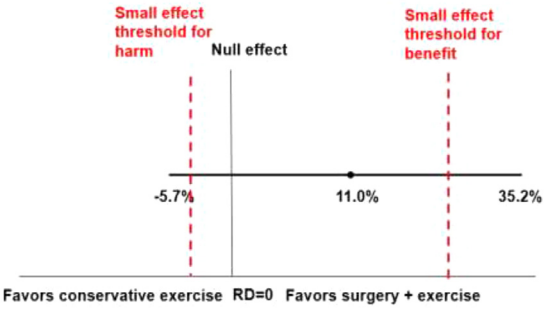


Appendix 3.2. Figure 2 Determining the target of certainty rating when results suggest an increase in harmful outcome (Example 2).

the body of evidence, the review team would have high certainty that the harm is small.

Example 3: When the 95% CI is extremely wide, reviewers could abandon being explicit about the target of certainty of evidence rating

Consider a systematic review of arthroscopic surgery plus postoperative exercise therapy versus conservative exercise therapy in adults with subacromial pain syndrome (SAPS) [3]. The authors found one RCT that addressed this issue. The RCT revealed that surgery plus exercise yielded 11 more patients achieving success per 100 patients (defined as "no shoulder problems at all"/ "healed completely" or "much better"), with a 95% CI from 5.7 fewer to 35.2 more per 100 patients at 6 months compared with conservative exercise (Appendix 3 Fig. 3.3). One might well judge that the 95% CI crosses both the threshold of large benefit and the threshold of small but important harm. Considering the width of the 95% CI, the systematic review team were very uncertain whether surgery improves or worsens global perceived effect at 6 months [3]. In such situations, there would be no need for authors to decide about the target for certainty rating, and they would rate down two levels for imprecision.



Appendix 3.3. Figure 3 When the 95% CI is extremely wide, reviewers could abandon being explicit about the target of certainty rating (Example 3).

References

1.Brignardello-Petersen R, Guyatt GH, Buchbinder R, *et al*, Knee arthroscopy versus conservative management in patients with degenerative knee disease: a systematic review. BMJ Open 2017;7:e016114.

2.Devji T, Guyatt GH, Lytvyn L, *et al*. Application of minimal important differences in degenerative knee disease outcomes: a systematic review and case study to inform BMJ Rapid Recommendations. BMJ Open 2017;7:e015587.

3.Hao QK, Tampi M, O'Donnell M, *et al*. Clopidogrel plus aspirin versus aspirin alone for acute minor ischaemic stroke or high risk transient ischaemic attack: systematic review and meta-analysis. BMJ 2018;363:k5108.

4.Lähdeoja T, Karjalainen T, Jokihaara J, *et al*. Subacromial decompression surgery for adults with shoulder pain: a systematic review with meta-analysis. British Journal of Sports Medicine 2020;54:665–673

References

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

[2] Noordzij M, van Diepen M, Caskey FC, Jager KJ. Relative risk versus absolute risk: one cannot be interpreted without the other. Nephrol Dial Transplant 2017;32:ii13–18.

[3] Schünemann JH. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? J Clin Epidemiol 2016;75:6–15.

[4] Brian S Alper BS, Oettgen P, Kunnamo I, Iorio A, Ansari MT, Murad MH, et al. Defining certainty of net benefit: a GRADE concept paper. BMJ Open 2019;9:e027445.

[5] Newcombe RG. Propagating imprecision: combining confidence intervals from independent sources. Commun Stat 2011;17:3154–80.

[6] Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol 2013;66:173–83.

[7] Schünemann HJ, Vist GE, Higgins JPT, Santesso N, Deeks JJ, Glasziou P, et al. Chapter 15: Interpreting results and drawing conclusions. Cochrane Handbook. Available at: https://training.cochrane.org/handbook/current/chapter-15#section-15-4-4-3 2021.

[8] Tikkinen KAO, Craigie S, Schünemann HJ, Guyatt GH, et al. Certainty ranges facilitated explicit and transparent judgments regarding evidence credibility. J Clin Epidem 2018;104:46–51.

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