OPEN ACCESS

Rating certainty when the target threshold is the null and the point estimate is close to the null

Linan Zeng ,[1,2,3,4] Monica Hultcrantz,[5,6] David Tovey,[7] Nancy Santesso,[3] Philipp Dahm ,[8] Romina Brignardello-Petersen ,[3,4] Reem A Mustafa,[9] M Hassan Murad ,[10] Ariel Izcovich ,[11] Hans de Beer,[12] Martin Alberto Ragusa,[13] Bradley Johnston,[14] Lingli Zhang,[1,2] Alfonso Iorio,[3,15] Gordon Guyatt[3,4,15]

10.1136/bmjebm-2024-113077

For numbered affiliations see end of article.

Correspondence to:
Dr Gordon Guyatt; guyatt@mcmaster.ca

Abstract

When one initially targets the null effect and the point estimate falls close to the null, two challenges exist in rating certainty of evidence. First, when the point estimate is near the null and the data, therefore, suggests little or no effect, rating certainty in a benefit or harm is misleading. Second, since in general the narrower the confidence interval (CI) the more precise the estimate, if the CI is narrow, rating down for imprecision due simply to crossing the null is inappropriate. This paper addresses these issues and provides a solution: to revise the target of certainty rating from a non-zero effect to a little or no effect. This solution requires estimating a range in which the minimal important difference (MID) for benefit and an MID for harm might lie, and thus establishing a range that represents little or no effect. If GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) users are confident that the point estimate represents an effect less than the smallest plausible MID, they will revise their target and rate certainty to a little or no effect. If the entire CI falls within the range of little or no effect, they will not rate down for imprecision. Otherwise (if the CI includes an important effect), they will rate down. Using the solution provided in this paper GRADE users can make an optimal choice of the target of certainty rating.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ In 2017, GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) published a concept paper in which authors clarified what it is in which those using GRADE are rating their certainty—the target of certainty rating. The relative location of a point estimate in relation to a target threshold(s) determines the target of certainty rating. Possible target thresholds include the null and the minimally important difference, and ranges include little or no effect, small, moderate and large effects.

WHAT THIS STUDY ADDS

⇒ Building on prior GRADE clarification, this paper first clearly identifies two challenges in situations in which GRADE users initially target the null and the point estimate is close to the null. The first changllenge is that rating certainty in a non-zero effect when the correct inference is that the effect is little or no effect is misleading. The second is rating down for imprecision when a narrow CI crosses the null is inappropriate.

⇒ This paper provides detailed guidance previously unavailable for GRADE users who initially choose to rate their certainty with respect to the null and observe a point estimate close to the null.

Check for updates

To cite: Zeng L, Hultcrantz M, Tovey D, *et al*. *BMJ Evidence-Based Medicine* 2025;**30**:202–207.

Introduction

GRADE guidance thus far

In 2017, GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) published a key paper in which the authors clarified what it is in which those using GRADE are rating their certainty—the target of certainty rating.[1] Previous to that paper, GRADE had specified the target as certainty in the point estimate. The 2017 paper pointed out the shortcomings of this conceptualisation and suggested an alternative that has since then become core GRADE guidance: GRADE users are rating their certainty that the true effect lies on one side of a threshold or in a particular range.[1]

Possible thresholds include the null and the minimally important difference (MID, also called the small effect threshold, defined as the smallest change in an outcome that people perceive as important).[2] Ranges can include little or no effect,

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This article alerts GRADE users to the challenges in determining the target of certainty rating when the point estimate is close to the null. GRADE users can consider applying the solution in this article to avoid giving a misleading certainty rating to their audience.

small, moderate and large effects.[1] One can also rate the certainty in a net benefit. In a paper published in 2021, GRADE offered additional guidance on deciding the target of the certainty rating, clarifying that the key determinant of the choice is where the point estimate lies in relation to the chosen threshold.[3]

Following current GRADE guidance, optimal GRADE use requires identifying the chosen threshold or range and rating the certainty accordingly.[1 3] If GRADE users believe it is most useful to their intended audience, they can choose the MID. If, on the other hand, they believe it is most relevant to their audience, and if in addition, they prefer to minimise making the value judgements that are required with the value-laden choices of a threshold, they can choose the null effect as the threshold of interest.

In this paper, we discuss challenges that arise when the initial target threshold is the null and the point estimate falls close to the null, and provide a potential solution to these challenges from which GRADE users can choose.

Two major challenges in certainty rating when the point estimate is close to the null and the CI crosses the threshold

Consider a hypothetical systematic review of intervention X for patients at risk of myocardial infarction. The review authors choose the null effect as the target threshold. For situation A1 in figure 1, when the authors rate their certainty that intervention X reduces the risk of myocardial infarction (ie, a non-zero effect, also called a non-null effect, exists), they will, because the confidence interval (CI) does not overlap the null, not rate down for imprecision. For situation A2, the authors can still rate their certainty that the intervention reduces the risk of myocardial infarction, but because the CI overlaps the threshold of interest, they will rate down for imprecision. These two scenarios are uncontroversial and can be addressed using existing GRADE guidance.

For situation B1, the point estimate becomes close to the null effect. As the point estimate still falls on the left side of the null the authors might continue to rate their certainty in a non-zero effect (in this case, intervention X reduces myocardial infarction). The choice would, however, be inappropriate because the correct inference is that the effect is a little or no effect. For situation B2, if the authors continue to rate their certainty in a non-zero effect, because the CI crosses the null, the authors would rate down for imprecision. That would, however, lead to a second reason for being inappropriate: the very narrow CI precludes rating down for imprecision.

Solution to the challenges: revising the target of certainty rating from a non-zero effect to a little or no effect

For situations B1 and B2 in figure 1, recognising that the null may indeed represent the true effect or if not the null, at least a value very near the null that represents little or no effect, review authors can revise their target of certainty rating from a non-zero effect (choosing the null as the target threshold) to a little or no effect (choosing a range of little or no effect as the target range).

Consider a real systematic review of lower blood pressure target (≤135/85 mm Hg) versus standard blood pressure target (≤140– 160/90–100 mm Hg) for patients diagnosed with cardiovascular disease and with high blood pressure.[4] A meta-analysis of seven randomised controlled trials (RCTs) including 9595 patients reported a point estimate of 0 fewer deaths per 1000 patients, with a CI from 10 fewer to 10 more deaths per 1000 patients (figure 2).

While the review authors did not explicitly specify their target of certainty rating, we consider that they started with the intention of rating certainty with a non-zero effect. As the point estimate falls very close to the null, rating certainty in either benefit

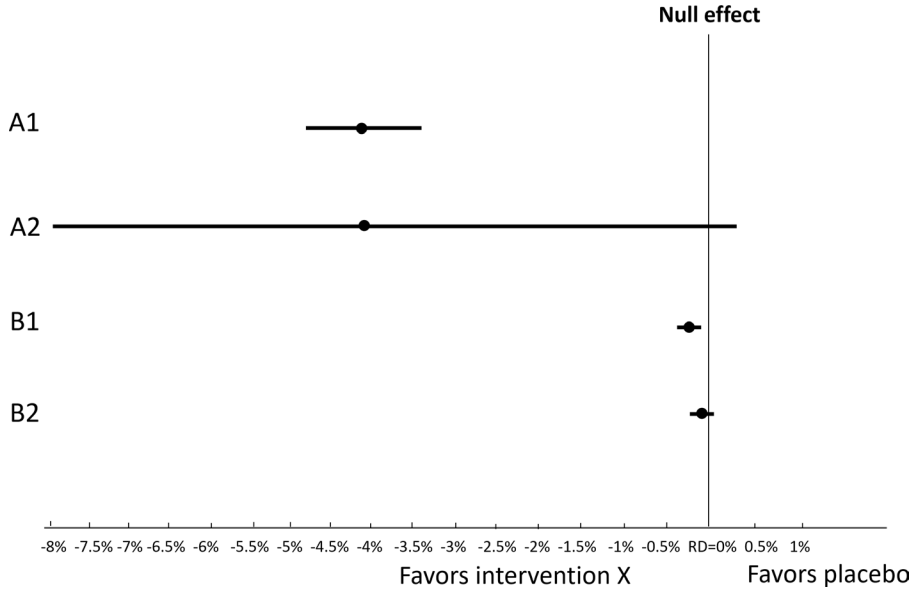

Figure 1 Two challenges arise when initially target the null and the point estimate turns out to be very close to the null: first, rating certainty in a non-zero effect when the point estimate indicates a little or no effect (situation B1); second, rating down for imprecision when the CI is very narrow (situation B2). RD, risk difference.

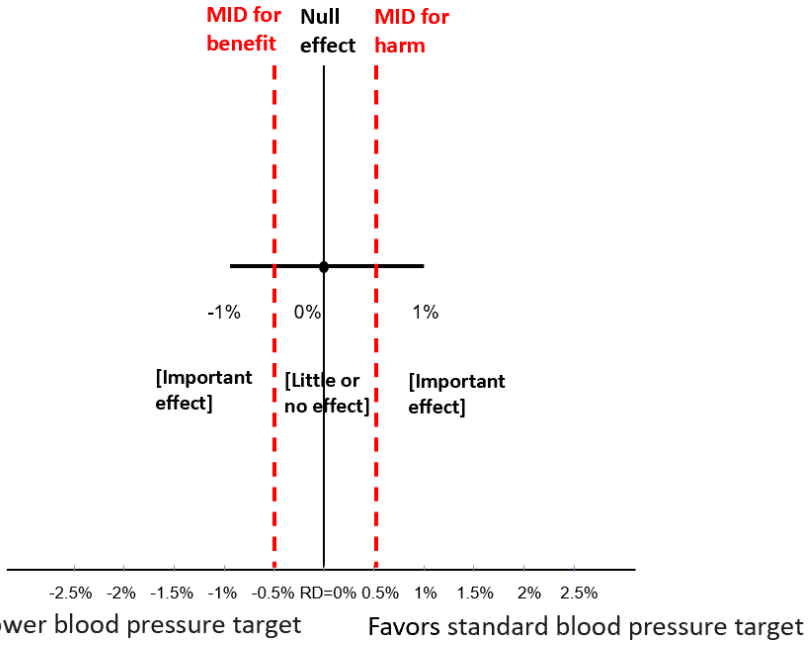

Figure 2 Meta-analysis of lower versus standard blood pressure target on total mortality in patients with cardiovascular disease and high blood pressure. MID, minimally important difference.

or harm becomes inappropriate and potentially misleading. Thus, revising the target of certainty rating to a certainty with a little or no effect is advisable.

When to revise the target of certainty of evidence rating

One might ask how close the point estimate needs to be to the null before one should consider revising the target of certainty rating from a non-zero effect to a little or no effect. The previous GRADE guidance does not provide an answer. We suggest that, once GRADE users conclude that the point estimate may represent little or no effect (ie, the point estimate is consistent with an effect less than the MID), they should consider revising the target of certainty rating. GRADE users often choose the null as the threshold to minimise value and preference judgements; when the point estimate is near the null such judgement becomes necessary.

Formal approaches to help establish MIDs are now available.[5–8] Alternatively, searching the literature for studies of values and preferences and health state utilities may be helpful. If the outcome is a patient-reported measurement instrument, such as a Visual Analogue Scale (VAS) score for pain intensity, GRADE users are likely to find relevant literature establishing a suggested MID.[9] Given the usual paucity of evidence regarding patients' values and the variation in patients' values, uncertainty regarding the MID is invariably appropriate. Thus, GRADE users may consider a range of plausible MIDs from the largest to the smallest plausible MID (figure 3).

If the point estimate is consistent with an increase in the effect that exceeds the largest plausible MID (A in figure 3), they would confidently rate certainty in a non-zero effect. If, on the other hand, the point estimate falls below the value they have designated as the smallest plausible MID, and the point estimate, therefore, represents little or no effect (C in figure 3), they would confidently revise the target of certainty rating to a little or no effect.

If they consider that the point estimate falls in the range of uncertainty regarding the MID (ie, they are uncertain if the point estimate is above or below some true MID), particularly if the point estimate is smaller than their best estimate of the MID (B in figure 3), they may or may not revise the target of certainty rating to a little or no effect. Either option would be reasonable as long as the authors transparently explain the underlying rationale.

For systematic review authors, exactly specifying the best estimate of MID or the exact boundaries of the range of plausible MIDs is often challenging. It turns out it may also be unnecessary. All review authors need to judge whether their specific point estimate is (1) clearly greater than the MID (they can then still rate certainty in a non-zero effect), (2) clearly smaller than the MID (they then revise to rate certainty with a little or no effect) or (3) the point estimate falls in the range of uncertainty around the MID (in which case they can either revise or not revise the target of certainty rating). Even if they choose to avoid the exact specification of an MID and an exact plausible range, understanding the logic underlying the choice of revising or not revising is helpful.

Imprecision rating after revising the target of certainty rating

After revising the target of certainty rating, making the imprecision judgement requires deciding whether the CI overlaps either of the two thresholds (ie, the MID for benefit and the MID for harm) that form the range of little or no effect. That is, when rating certainty in a little or no effect, GRADE users judge whether both ends of the CI still represent a little or no effect.

As discussed above GRADE users can acknowledge the uncertainty or arbitrariness of MID and potentially comment on how reasonable alternative choices would impact decisions regarding rating down for imprecision.

In the systematic review of lower versus standard blood pressure target (figure 2), had the authors set the MID of mortality as a difference around 5 per 1000, as the CI overlaps with both boundaries of the range (ie, the CI includes both important benefit and important harm), they would confidently rate down at least once for imprecision and may rate down twice.[10] If the authors had no concerns regarding the other four GRADE domains, the

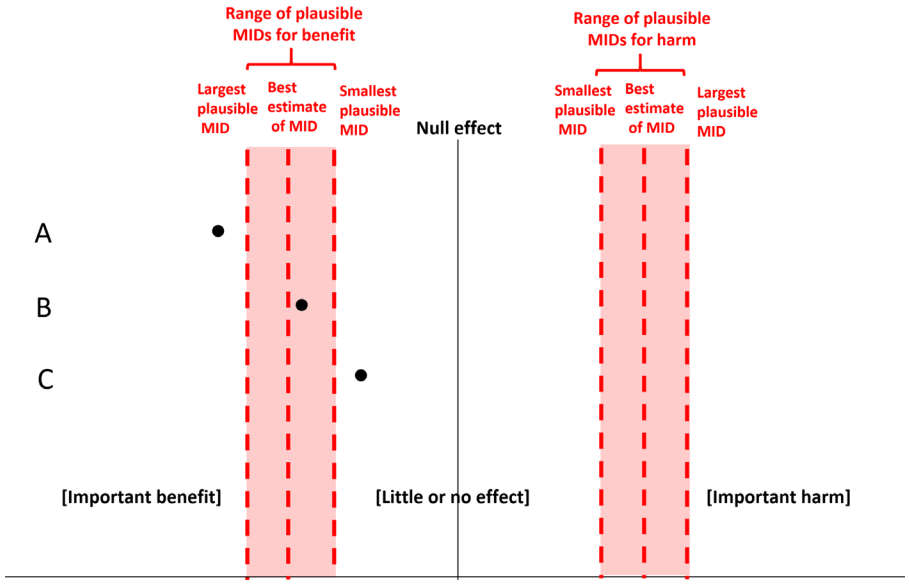

Figure 3 The relative location of the point estimate to a range of plausible MIDs. In situation A, the point estimate is consistent with an increase in the effect that exceeds the largest plausible MID, and GRADE users can confidently rate certainty in a non-zero effect. In situation C, the point estimate falls below the value they have designated as the smallest plausible MID, and they can revise the target of certainty rating to a little or no effect. In situation B, the point estimate falls in the range of uncertainty regarding the MID, they may or may not revise the target of certainty rating to a little or no effect. GRADE (Grading of Recommendations, Assessment, Development, and Evaluations). MID, minimally important difference.

plain language summary would be that the lower blood pressure target likely has (rating down once for imprecision) or may have (rating down twice for imprecision) little or no effect on mortality. Table 1 illustrates the GRADE plain language summary for each level of certainty of evidence.

Another two examples of deciding on the target of certainty rating when the point estimate is close to the null

Consider a systematic review addressing corticosteroids versus no corticosteroids in patients with sepsis.[11] A meta-analysis of 17 RCTs with 4243 participants reported a point estimate of 3 more gastrointestinal bleeds per 1000 patients in patients randomised to corticosteroids, with a CI from 5 fewer to 13 more per 1000 patients.

Had review authors initially planned to rate certainty in a non-zero effect, as the point estimate turned out to be close to the null effect, they would consider whether they should revise the target of the certainty rating. Without considering what the exact value of MID might be, they might reasonably conclude that an increase of 3 in 1000 is clearly smaller than the MID. Having made this inference, they would revise the certainty target to a little or no effect and move to consider the boundaries of the CI. If their judgement is that a 13 per 1000 increase also represents little or no effect, as does a 5 per 1000 decrease (ie, both ends of the CI fall within the range of little or no effect), they would not rate down for imprecision. They can do so without deciding at what value above 13 per 1000 the MID lies. Assuming the authors have no concerns for the other four GRADE domains (ie, risk of bias,inconsistency, indirectness, publication bias), the plain language summary would state that corticosteroids have little or no effect on gastrointestinal bleeding.

The solution we suggest for considering the target of certainty rating also applies to continuous outcomes. Consider a systematic review of closed versus open kinetic chain exercises for patients with patellofemoral pain syndrome.[12] A meta-analysis of three RCTs including 122 patients reported a point estimate of an increase of 0.03 points on a 0–10 VAS (visual analogue scale) with a CI from a decrease of 0.37 to an increase of 0.76 points. The empirical evidence suggests that the MIDs on a 0–10 VAS range from one to two points.[13 14]

Had the review authors initially planned to rate certainty in a non-zero effect, as the point estimate is clearly smaller than the MID, they would revise the target of certainty rating from null effect to a little or no effect. As the entire CI falls within the range of little or no effect, the authors would not rate down for imprecision. Given the review authors have no concerns on the remaining four GRADE domains, the plain language summary would be closed kinetic chain exercises, compared with open kinetic chain exercises, have little or no effect on pain intensity.

Table 1 GRADE (Grading of Recommendations, Assessment, Development, and Evaluations) plain language summary

Certainty	Plain language summary
High	'Treatment has an(a) (little or no/important/small/moderate/large) effect.'
Moderate	'Treatment likely/probably has an(a) (little or no/important/small/moderate/large) effect.'
Low	'Treatment may have an(a) (little or nol/important/small/moderate/large) effect.'
Very low	'We are very uncertain about the effect of the treatment.'

Discussion

This article has, using hypothetical and actual examples, introduced the challenges that arise when authors initially decide to rate certainty in a non-zero effect and the point estimate falls close to the null. Our suggested approach considers whether the point estimate is certainly or possibly less than the MID: if certainly so (ie, clearly smaller than the smallest plausible MID), we suggest revising to rate certainty in a little or no effect. If possibly so (ie, falls within the range of plausible MIDs, particularly below the best estimate of MID), options of continuing to rate certainty in a non-zero effect or revising remain.

In deciding between the null and MID as the target of the certainty rating systematic review authors are likely to choose the null when they anticipate non-zero effects of an intervention. They are likely to choose the MID when they anticipate that differences between candidate interventions will represent little or no effect.

The solution of considering switching the target of certainty rating from a non-zero effect to a little or no effect when the point estimate turns out to be close to the null is applicable in both the context of systematic reviews and clinical guidelines. That is, while it is far less likely that guideline developers will choose the null as the target of certainty rating than systematic reviewers will do so, once they have chosen this threshold our guidance will apply.

In a systematic review or guideline in which authors have chosen to rate the certainty in a non-zero effect, their protocol should acknowledge the possibility that point estimates may be near the null and acknowledge that in such case revising the target of the certainty rating to either a little or no effect may be necessary. They can do this easily by pointing out that revising the target may occur and, if it does, they will follow the solution in this paper. As defining 'near the null' leads to considering MID, they may then add (or not add) additional details. Authors concerned about the specification of thresholds after rather than before summarising the evidence will add these additional details, those less concerned will not.

The solution we provide in this paper might not be the only reasonable approach for determining the target of certainty rating in situations in which the point estimate is close to the initially chosen threshold. For example, in the discussions within our co-authors' group, another two possible solutions raised. The first is to consider that any threshold is not exact but a range. When the point estimate is close to the target threshold, authors can rate their certainty in relation to the range that they consider representing the plausible thresholds. The second is to consider switching to rate certainty in relation to a prespecified 'indifference margin' (ie, similar to a non-inferiority threshold in a non-inferiority trial). If the point estimate falls within the range that formed by the indifference margin for benefit and the difference margin for harm, the authors could rate certainty that the intervention is non-different to the comparison.

Conclusions

This GRADE-related article has introduced the challenges that occur when GRADE users initially decide to rate the certainty in a non-zero effect and the point estimate falls close to the null. GRADE users should note these challenges and can consider applying the solution we suggest to transparently determine the target of their certainty rating and to make imprecision judgement accordingly.

Author affiliations
[1]Pharmacy Department/Evidence-based Pharmacy Centre/Children's Medicine Key Laboratory of Sichuan Province, West China Second University Hospital, Sichuan University; Sichuan University and Key Laboratory of Birth Defects and Related Disease of Women and Children, Ministry of Education, West China Second University Hospital, Chengdu, People's Republic of China
[2]West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China
[3]Department of Health Research Methods, Evidence, and Impact, McMaster University, Ontario, Canada, McMaster University, Hamilton, Ontario, Canada
[4]MAGIC Evidence Ecosystem Foundation, Oslo, Norway, MAGIC Evidence Ecosystem Foundation, Oslo, Norway
[5]HTA Region Stockholm, Solna, Stockholm County, Sweden
[6]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden
[7]Journal of Clinical Epidemiology, Sussex, UK
[8]Minneapolis Veterans Healthcare System, Minnepolis, Minnesota, USA
[9]Division of Nephrology and Hypertension, Department of Internal Medicine, University of Kansas Medical Centre, Kansas City, USA, University of Kansas Medical Centre, Kansas City, Kansas, USA
[10]Evidence-based Practice Center, Mayo Clinic, Rochester, Minnesota, USA
[11]Medicine Department, Universidad del Salvador, Buenos Aires, Argentina
[12]Guide2Guidance, Utrecht, Netherlands
[13]Internal Medicine Division, Fernandez Hospital, Buenos Aires, Argentina
[14]Department of Nutrition, College of Agriculture and Life Sciences:Texas A&M University; Department of Epidemiology & Biostatistics, School of Public Health:Texas A&M University, Texas A&M University, College Station, Texas, USA
[15]Department of Medicine, McMaster University, Ontario, Canada, McMaster University, Hamilton, Ontario, Canada

X Philipp Dahm @EBMUrology and Ariel Izcovich @IzcovichA

Acknowledgements The initial version of this paper was included in a thesis, LZe. Methodological Issues in Rating Certainty of Evidence and Interpreting Magnitude of Effect in Systematic Reviews and Practice Guidelines (Available at: https://macsphere.mcmaster.ca/handle/11375/29554).

Contributors GG, LZe, MH and DT contributed to the conception of the project. LZe and GG drafted the manuscript with all other authors having revised it critically for important intellectual content. GG is the guarantor. All authors have approved the final manuscript. All authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding This project was supported by the National Natural Science Foundation of China (72474148) and the Einstein Foundation Berlin as part of the Einstein Foundation Award for Promoting Quality in Research.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iDs
Linan Zeng http://orcid.org/0009-0009-5081-3353

Philipp Dahm http://orcid.org/0000-0003-2819-2553
Romina Brignardello-Petersen http://orcid.org/0000-0002-6010-9900
M Hassan Murad http://orcid.org/0000-0001-5502-5975
Ariel Izcovich http://orcid.org/0000-0001-9053-4396

References

1 **Hultcrantz M**, Rind D, Akl EA, *et al*. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol* 2017;87:4–13.
2 **Schünemann HJ**. Hello MID, where do you come from. *Health Serv Res* 2005;40:593–7.
3 **Zeng L**, Brignardello-Petersen R, Hultcrantz M, *et al*. GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings. *J Clin Epidemiol* 2021;137:163–75.
4 **Saiz LC**, Gorricho J, Garjón J, *et al*. Blood pressure targets for the treatment of people with hypertension and cardiovascular disease. *Cochrane Database Syst Rev* 2022;11:CD010315.
5 **Morgano GP**, Mbuagbaw L, Santesso N, *et al*. Defining decision thresholds for judgments on health benefits and harms using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks: a protocol for a randomised methodological study (GRADE-THRESHOLD). *BMJ Open* 2022;12:e053246.
6 **Zeng L**, Helsingen LM, Bretthauer M, *et al*. A novel framework for incorporating patient values and preferences in making guideline recommendations: guideline panel surveys. *J Clin Epidemiol* 2023;161:164–72.
7 **Helsingen LM**, Zeng L, Siemieniuk RA, *et al*. Establishing thresholds for important benefits considering the harms of screening interventions. *BMJ Open* 2020;10:e037854.
8 **Neumann I**, Quiñelen E, Nahuelhual P, *et al*. Using Explicit Thresholds were valuable for judging Benefits and Harms in partially contextualized GRADE Guidelines. *J Clin Epidemiol* 2022;147:69–75.
9 **Devji T**, Carrasco-Labra A, Guyatt G. Mind the methods of determining minimal important differences: three critical issues to consider. *Evid Based Ment Health* 2021;24:77–81.
10 **Zeng L**, Brignardello-Petersen R, Hultcrantz M, *et al*. GRADE Guidance 34: update on rating imprecision using a minimally contextualized approach. *J Clin Epidemiol* 2022;150:216–24.
11 **Rochwerg B**, Oczkowski SJ, Siemieniuk RAC, *et al*. Corticosteroids in Sepsis: An Updated Systematic Review and Meta-Analysis. *Crit Care Med* 2018;46:1411–20.
12 **van der Heijden RA**, Lankhorst NE, van Linschoten R, *et al*. Exercise for treating patellofemoral pain syndrome. *Cochrane Database Syst Rev* 2015;1:CD010387.
13 **Kelly AM**. The minimum clinically significant difference in visual analogue scale pain score does not differ with severity of pain. *Emerg Med J* 2001;18:205–7.
14 **Olsen MF**, Bjerre E, Hansen MD, *et al*. Pain relief that matters to patients: systematic review of empirical studies assessing the minimum clinically important difference in acute pain. *BMC Med* 2017;15:35.