Core GRADE 3: rating certainty of evidence—assessing inconsistency

Gordon Guyatt,[1,2,3] Stefan Schandelmaier,[4,5,6] Romina Brignardello-Petersen,[1] Hans De Beer,[7] Manya Prasad,[8] M Hassan Murad,[9] Prashanti Eachempati,[3,10,11] Derek K Chu,[1,2] Rohan D'Souza,[1,12] Alfonso Iorio,[1,2] Thomas Agoritsas,[1,3,13] Liang Yao,[14] Reem A Mustafa,[1,15] Sameer Parpia,[1] Pasqualina Santaguida,[1] Per Olav Vandvik,[3,16] Monica Hultcrantz,[17,18] Victor M Montori[19,20]

For numbered affiliations see end of the article

Correspondence to: G Guyatt guyatt@mcmaster.ca; (ORCID 0000-0003-2352-5718)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* **2025;389:e081905** http://dx.doi.org/10.1136/bmj-2024-081905

Accepted: 14 March 2025

This third article in a seven part series presents the Core GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach to deciding whether to rate down certainty of evidence due to inconsistency—that is, unexplained variability in results across studies. For binary outcomes in which relative effects are consistent across baseline risks while absolute effects are not, Core Grade users assess consistency in relative effects. For continuous outcomes, they assess consistency in the absolute effects. When planning for the possibility of inconsistent results across studies, systematic review authors using Core GRADE construct a priori hypotheses regarding population or intervention characteristics that may explain inconsistency. They then judge the magnitude of inconsistency by considering the extent to which point estimates differ and the degree to which confidence intervals overlap. Before making a decision on rating down, Core GRADE users will evaluate where individual study estimates lie in relation to the threshold of the certainty rating (minimal important difference or the null). Finally, they will test their subgroup hypothesis and if an effect proves credible will provide separate evidence summaries and rate certainty of evidence separately for each subgroup. When they find no credible subgroup effect, they will provide a single evidence summary, rating down for inconsistency if necessary.

SUMMARY POINTS

- Inconsistency refers to unexplained variability in results across studies
- For binary outcomes, Core GRADE (Grading of Recommendations Assessment, Development and Evaluation) focuses on the consistency of relative effects (eg, risk ratios or odds ratios)
- To address rating down for inconsistency, Core GRADE relies on the visual inspection of forest plots for the magnitude of differences in point estimates, the overlap of confidence intervals, and the relation of study estimates to the chosen threshold of the null effect or minimal important difference
- Higher values for the I2 statistic indicate greater inconsistency but may be misleading; thus Core GRADE users should interpret the I2 statistic cautiously
- Key criteria for determining the credibility of a subgroup analysis include the P value associated with a test of interaction, consistency with a priori hypotheses that include direction of effect, and whether the subgroup effect is based on within study comparisons
- If a subgroup effect is judged credible and substantial, Core GRADE users will present estimates separately for the relevant subgroups

This is the third paper in a series describing Core GRADE (Grading of Recommendations Assessment, Development and Evaluation), the essentials of the GRADE approach to rating certainty of evidence and grading recommendations for paired interventions and comparators focusing on the perspective of patients and clinicians. The previous two papers provided an overview of the Core GRADE process,[1] what to consider when choosing the target of the certainty rating, and how issues of imprecision can influence certainty ratings of a body of evidence.[2] In this paper, we address issues of inconsistency.

By inconsistency we mean unexplained variability in results across studies. We are particularly concerned about inconsistency that is sufficiently great that, depending on which of the varying results represents the truth, inferences for clinical practice would differ. Authors writing about inconsistency sometimes use the term heterogeneity, particularly when referring to statistical tests related to inconsistency.[3]

To best address inconsistency, Core GRADE users must first understand the measure of effect to which they should attend. When dealing with binary outcomes they should focus on relative effects such as risk ratios or hazard ratios, and when dealing with continuous outcomes they should focus on absolute effects such as mean differences. Next, they must prepare for the possibility of encountering large inconsistency by making a priori hypotheses that might explain that inconsistency. They must then review the results, decide if problematic inconsistency exists, and determine if the a priori hypotheses they have generated explain the inconsistency. If after considering these hypotheses, large unexplained inconsistency remains, they will rate down the certainty of the evidence. This paper discusses each of these steps.

The information in this article will enable Core GRADE users to choose the appropriate effect measures for continuous and binary outcomes for judging inconsistency of evidence, choose appropriate a priori subgroup hypotheses to explain possible inconsistent results, judge inconsistency of evidence using appropriate visual and statistical criteria, and judge credibility of identified subgroup effects using a formal instrument.

Choosing the right measure of effect when assessing inconsistency

Binary outcomes: variability in relative versus absolute effects

As pointed out in the first paper in this series that provided an overview of the Core GRADE approach,[1] relative treatment effects seldom vary across patient subgroups such as old and young, male and female, or less sick and more sick.[4-9] However, given that such patient characteristics are often associated with substantial differences in baseline risk (ie, probability of experiencing the outcome in the comparator group), even in the presence of constant relative treatment effects across such patient groups, the resulting absolute treatment effects will differ substantially.

The hypothetical example in figure 1 illustrates the situation. Here, the relative risk reduction is constant, at 33%, across low, medium and high risk groups. Because of the substantial differences in baseline risk, the risk difference between treated and untreated patients varies substantially, from 10% in high risk patients to 1% in low risk patients.

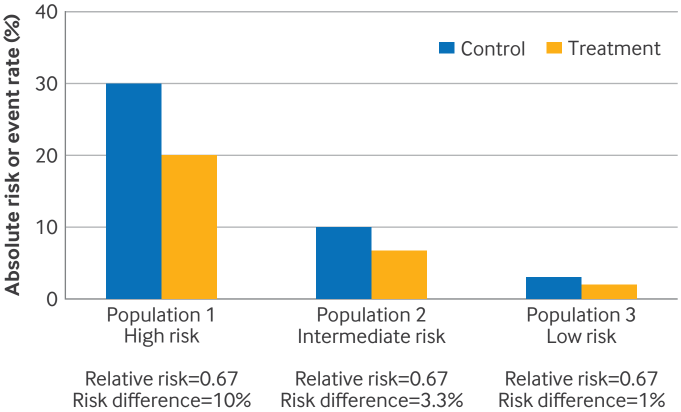

Fig 1 | Constant relative risk with varying baseline risk, leading to varying reduction in absolute risks. In each population, the larger event rate represents the control (baseline risk, blue bars) and the smaller event rate the intervention (orange bars)

Despite risk differences being more important to patients than relative risks, authors of randomised trials and meta-analyses typically highlight relative rather than absolute effects. They do so because of the typical consistency in relative risks and the expected variability in risk differences (see fig 1). Greater consistency in results is desirable: it increases confidence in the pooled estimates of effect. Thus, the anticipated consistency of relative effects and variability of absolute effects is the reason why, in Core GRADE summary of findings tables, risk differences are estimated in each relevant patient group by applying relative risks to baseline risks, and why guideline authors may offer different treatment recommendations for individuals at low, medium, and high risk. Finally, because inconsistency in absolute effects is ubiquitous and inconsistency in relative effects is rare, we are concerned with inconsistency in relative rather than absolute effects.

Continuous outcomes

Continuous outcomes are typically measured as absolute effects—thus, when considering inconsistency, looking at relative effects is typically not an option. For example, duration of illness, hospital length of stay, functional status, or quality of life are typically evaluated as mean differences. Inconsistency in mean differences across studies can lower certainty in evidence in the same way as inconsistency in relative effects does for binary outcomes.

Core GRADE's approach to preparing for inconsistency

In this section, we discuss how, when thinking ahead to possible inconsistency in results, Core GRADE users formulate a plan to best deal with the inconsistency they may ultimately find. In general terms, when observing relative effects for binary outcomes and absolute effects for continuous outcomes across studies in a body of evidence, there may be several reasons for inconsistency. These include random error and differences in population, intervention, comparison, and outcome (PICO) elements. Hypotheses may be able to explain these differences—this is the hope when preparing for the possibility of large inconsistency—or they may not. If they do explain inconsistency, Core GRADE users will provide separate evidence summaries for each subgroup and make judgments about inconsistency within each subgroup. If the hypotheses do not explain differences, the unexplained variability in effects decreases the certainty of evidence.

Variability in PICO elements

Core GRADE ratings of certainty pertain to bodies of evidence summarised in rigorous systematic reviews. The Core GRADE process begins with construction of a structured clinical question.[1] Studies addressing a particular question are certain to vary in patients enrolled, aspects of the intervention and comparator chosen, and the way the outcome is measured, and such variability is often appreciable.

Core GRADE users may intuit that such variability (ie, inconsistency in PICO elements) compromises the certainty of evidence from a systematic review. This, however, is rarely the case. Indeed, if effects are similar from study to study, variability in the PICO elements enhances the applicability of the pooled effect to a wider range of clinical contexts. If effects vary across studies, differences in the PICO elements provide an opportunity to explore the possible sources of the inconsistency in results. Thus, inconsistency in PICO elements is not what decreases confidence in the evidence.

Three options for possible subgroups with different intervention effects

When reflecting on the possibility that effects differ across patient subgroups (eg, effects may differ in old versus young people) or across interventions subgroups (eg, oral versus parenteral antibiotic treatment), review authors face a potential problem. Selecting a narrow range of subgroups in the PICO will always sacrifice applicability, and often precision. Selecting a broader range of patient and intervention subgroups will enhance generalisability and precision but runs the risk, if effects differ substantially, of pooling inappropriately across patient or interventions subgroups.

To solve the problem Core GRADE users must, for each subgroup, distinguish between three scenarios: one has no reason to suspect differences in effects across subgroups; one is confident that effects vary across subgroups; or one has good reason to suspect subgroup differences but is uncertain.

Take, for example, two different age groups: young and old. The following are the three scenarios and corresponding actions they would mandate:

1. Previous research provides little support for the possibility that effects differ between old and young people. In this scenario, review authors would choose a broad age range for the PICO, and the findings would apply to both age groups.
2. Previous research has given reason to be confident that the relative effects on older versus younger people differ. Accordingly, one would choose a narrow age range for the PICO (eg, older people) or create two separate PICOs and sets of recommendations, one for older people and the other for younger people.
3. Previous research plausibly suggests that effects differ between old and young people, but one is uncertain. One would then choose a broad age range in the PICO and conduct subgroup analysis or meta-regression to explore the possible impact of differences in age.

Table 1 summarises the three scenarios when considering subgroups during PICO construction, and provides examples of each.

We recommend that to maximise precision and generalisability, review authors frame their PICOs broadly. In doing so, however, they must prepare themselves for the possibility of inconsistent results across studies. One way to prepare is to choose the third scenario when constructing the PICO. We now present details of how to deal with this third scenario.

Need for a priori hypotheses with a specified direction

As pointed out in the first article in this series,[1] preparation for the possibility of inconsistency in results involves generating a small number of well chosen a priori hypotheses to explain that inconsistency. Subgroup effects exist when the effects of an intervention versus a comparator differ according to characteristics of patients (eg, older versus younger, more sick versus less sick) or differences in interventions (eg, longer versus shorter duration of therapy). Thus, authors may postulate subgroup effects according to different patient groups or interventions.

These hypotheses should be based on previous evidence (eg, from a related trial, meta-analysis, or cohort study) or thorough understanding of the underlying biology, and they should include the direction of the subgroup effect (hypothesising, for example, not just that effects may differ across patient ages, but also that effects will be larger in old people than in young people). Postulating more than a small number (ideally three or fewer) of directional hypotheses will increase the likelihood of chance

Table 1 | Three scenarios when considering subgroups during PICO construction

Scenario	Implications for PICO construction	Example
1. Previous research provides no compelling evidence that effects differ across patient or intervention subgroups (no subgroup hypothesis)	Combine all subgroups (single estimate of effect) without a subgroup hypothesis	The World Health Organization has generated several recommendations regarding the management of patients with covid-19. The guideline panels inferred that effects were very likely to be similar in men and women and thus in all their recommendations provided a single estimate for men and women[10]
2. Previous research suggests that effects differ across patient or intervention subgroups (subgroup effects are presumed to exist)	Narrow PICO to one subgroup, or construct two separate PICOs for each subgroup	A guideline panel addressing opimal transfusion thresholds in anaemic patients considered that the biology differed between children and adults and therefore looked at the evidence separately and provided separate recommendations[11]
3. Previous research plausibly suggests that effects differ across patient or intervention subgroups, but one is uncertain (directional subgroup hypothesis)	Initially combine all subgroups (single estimate of effect), but also provide and then test a directional subgroup hypothesis	A systematic review comparing immediate versus delayed antiretroviral therapy in patients with a concomitant diagnosis of HIV and tuberculosis tested whether the impact of early versus delayed treatment on mortality differed between those with higher and lower CD4 cell counts.[12] A previous trial suggested that hypothesis, including a clear direction, but for another outome[13]

PICO=population, intervention, comparison, and outcome.

findings (spurious associations), thus undermining the credibility of any subgroup effects.

For instance, in the systematic review of when to start antiretroviral therapy in patients with a concomitant diagnosis of tuberculosis and HIV, the authors made only a single a priori hypothesis in considering mortality. They postulated that effects may differ depending on CD4 T cell counts using a threshold of $<0.050\times10^9$ cells/L *v* $>0.050\times10^9$ cells/L.[11,12] Their hypothesis was based on previous evidence of a higher incidence of adverse immune reactions in patients with a lower CD4 T cell count.[13] One might reasonably presume the direction of the subgroup effect (early antiretroviral therapy is worse in those with lower CD4 T cell counts). In this case, as it turned out, and contrary to the hypothesis, the results suggested that if there was a benefit of early therapy it was more likely in those with a low CD4 T cell count (P=0.12 for interaction). The example thus highlights how the review authors prepared themselves for the possibility of inconsistent results through specifying a single, directed subgroup hypothesis based on related evidence. The example also highlights that, without a subgroup analysis, Core GRADE users should exercise caution before concluding that effects differ between subgroups.

The ability to predict the direction of a subgroup effect provides a useful criterion when deciding between the first scenario (broad PICO, no subgroup analysis) and third scenario (broad PICO and subgroup analysis) discussed earlier. If one cannot confidently specify the direction of the potential subgroup effect, one should choose the first scenario rather than the third. Consistent with our recommendation of a small number of compelling subgroup hypotheses, we discourage post hoc exploration of possible subgroup effects.

Criteria for judging serious inconsistency

Having addressed how Core GRADE users should plan for dealing with inconsistency in results, we will now address how they will implement their plan (see fig 2). In the three following sections we describe how Core GRADE users can determine whether inconsistency is of sufficient concern to consider rating down for inconsistency. If they do find important inconsistency, they should look to their a priori hypotheses to see if they can explain that inconsistency—a process that will include rating the credibility of any possible subgroup effects they identify. A subsequent section deals with this issue of subgroup explanations of variability in results. If only one eligible study exists, Core GRADE users will not rate down for inconsistency, although if the authors provide the data then they may still address the possibility of subgroup effects.

Three visual criteria from forest plots

Consider the hypothetical body of evidence in figure 3. When considering whether studies yield similar or different results, most observers of these forest plots will quickly conclude that results in the top half of the figure are consistent whereas results in the bottom half are inconsistent. Aspects of the results that justify these inferences are similarity versus differences in point estimates, the extent of overlap in confidence intervals (CIs), and the relation of point estimates to the threshold of certainty rating.

Point estimates—One is more inclined to consider rating down for inconsistency when point estimates differ substantially between studies. In figure 3, the point estimates in the top half of the figure are similar, ranging from 0.71 to 0.76. The similarity in the point estimates suggests no need to consider rating down for inconsistency. In contrast, in the bottom half of figure 3, two studies suggest substantial treatment effects—relative risk reductions >50%—and two other studies suggest modest harms, 17% and 25% increases in relative risk. The large differences in the point estimates of the two pairs of studies suggest rating down for inconsistency.

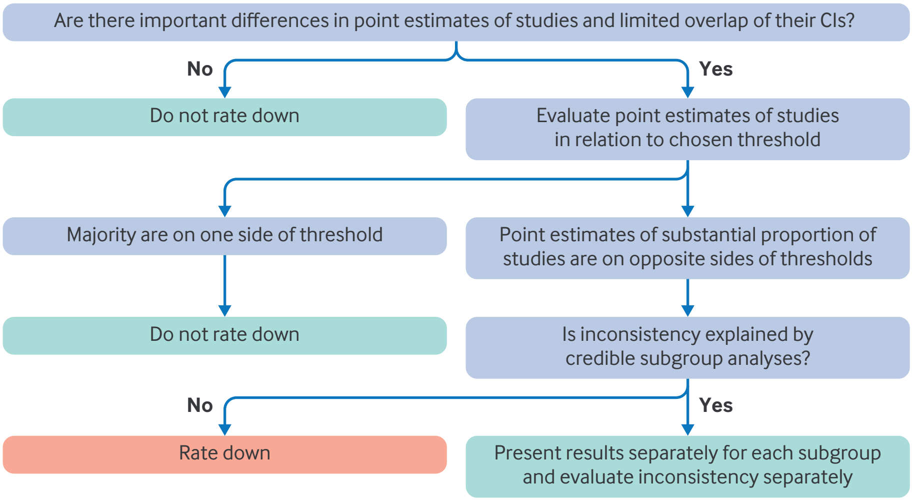
Are there important differences in point estimates of studies and limited overlap of their CIs?
No
Yes
Do not rate down
Evaluate point estimates of studies in relation to chosen threshold
Majority are on one side of threshold
Point estimates of substantial proportion of studies are on opposite sides of thresholds
Do not rate down
Is inconsistency explained by credible subgroup analyses?
No
Yes
Rate down
Present results separately for each subgroup and evaluate inconsistency separately

Fig 2 | Flow chart summarising Core GRADE's approach to addressing inconsistency in results. CI=confidence interval; GRADE=Grading of Recommendations Assessment, Development and Evaluation

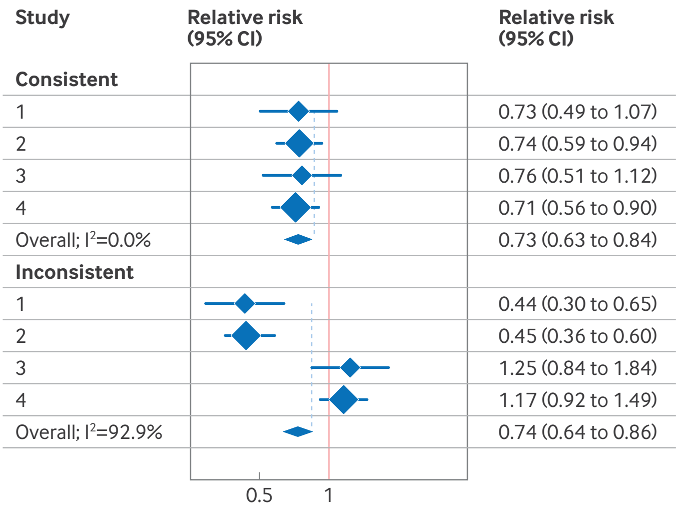

Fig 3 | Forest plots of consistent and inconsistent results from four randomised trials with similar overall pooled effects. The broken line represents the minimal important difference

Overlap of CIs—One is more inclined to rate down for inconsistency if the CIs of included studies do not show substantial overlap. In the top half of figure 3, the CIs of the four studies are largely overlapping. This overlap suggests no need to consider rating down for inconsistency. In contrast, in the bottom half of figure 3 the CIs between the first and second pairs of studies are completely non-overlapping. This provides a strong rationale for rating down for inconsistency.

Relation of point estimates to the threshold of certainty rating—Infrequently, Core GRADE users will find appreciable inconsistency using the first two criteria, but that point estimates largely lie on the same side of a chosen threshold (the null—ie, no difference between intervention and comparator) or the minimal important difference (MID)[2]). In these situations, they will be less inclined to rate down for inconsistency.

Whichever threshold one uses, in the top half of figure 3 all studies are on one side of the threshold (no need to consider rating down for inconsistency). In the bottom half of figure 3, the pairs of studies are on opposite sides of either threshold, with one pair showing benefit and the other showing harm, thus the need to consider rating down. In concluding important inconsistency, the lack of overlap of CIs is crucial.

While, as here, we may make initial assessments of inconsistency using relative risks, Core GRADE users must establish MIDs only on absolute risks. In this hypothetical example, the authors have, considering the baseline risk of the outcome, established that a relative risk reduction of about 15% will translate into a minimally important absolute effect of 1%. Supplementary appendix 1 describes this process.

Applying visual criteria: how choice of threshold affects judgments of inconsistency

The three key criteria for judging inconsistency—similarity of point estimates, overlapping of CIs, and relation of results to the chosen threshold for rating certainty—apply equally well to continuous outcomes. Consider figure 4, which depicts the results of a meta-analysis evaluating the impact of local infiltration analgesia on postoperative pain in patients after total knee arthroplasty (adapted from a figure we used in a previous GRADE article to illustrate these criteria).[14,15]

Consider the appropriate inference if authors of the systematic review of this evidence chose to rate their certainty with respect to the null. The pooled estimate clearly excludes the null, and the point estimates of all but one study support that inference. Thus, there is no reason to rate down for inconsistency.

However, consider if the review authors chose to rate their certainty with respect to the MID and selected a value of 10 mm. Now, five studies show values below the threshold and eight at or above the threshold. This inconsistency undermines the inference of an important effect suggested by the pooled estimate (14 mm) and would warrant rating down for inconsistency.

Although this example highlights how Core GRADE users should attend to the relation of point estimates to the threshold of certainty rating, when point estimates differ substantially and CIs do not overlap, they will seldom find compelling reason to invoke this additional criterion.

One criterion for statistical assessment and possible rating down twice for inconsistency

A statistical criterion, I^2, describes the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error (chance),[16] and may complement the three visual criteria. The lowest possible I^2, 0%, tells us that chance easily explains the difference between studies—the conclusion in the top half of figure 3. As I^2 approaches the highest possible value, 100%, the likelihood that chance alone explains the variability observed becomes extremely small. This is true of the bottom half of figure 3 in which I^2 is 93%.

I^2 may, however, prove misleading.[17-19] In particular, if the included studies have narrow CIs the associated I^2 may be misleadingly large. Moreover, if the point estimates are mostly on one side of the threshold of certainty rating, the high I^2 will be irrelevant. For instance, in figure 4 the high I^2 value of 95% suggests enormous inconsistency. Nevertheless, when using the null as the target of certainty ratings, 12 of the 13 studies showed mean differences favouring the intervention and one would conclude no problematic inconsistency.

It is natural that review authors desire hard and fast rules for interpreting I^2. The limitations of the statistic make such rules problematic. The best we can do is suggest that one will seldom see serious inconsistency with I^2 values <30%, and as I^2 rises beyond that value, the possible need to rate down certainty increases.

A final issue is consideration of rating down twice for inconsistency. Although this is a theoretical possibility, we have found compelling reason to rate down twice

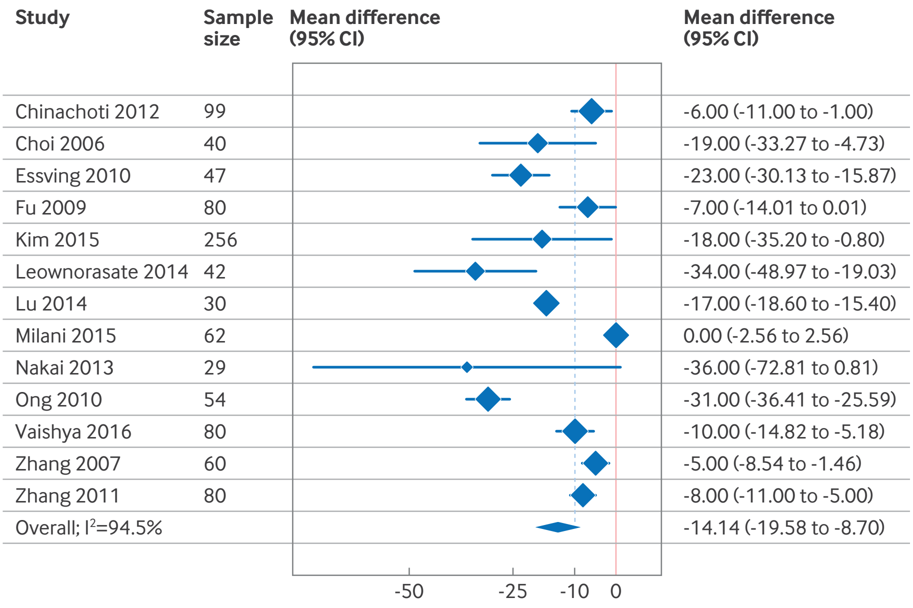

Fig 4 | Forest plot from a systematic review on the impact of local infiltration analgesia on postoperative pain after total knee arthroplasty. The broken line represents an estimate of the minimal important difference in pain score (10 mm) on a 100 mm visual analogue scale. Adapted from Guyatt et al[14]

for inconsistency sufficiently unusual that it need not concern users of Core GRADE.

Apparent subgroup effects based on a priori hypotheses

The burden of proof lies with those claiming a subgroup effect

We have pointed out that relative effects overwhelmingly tend to be similar across subgroups, and testing a large number of subgroup hypotheses results in a high risk of spurious findings. In general, Core GRADE users should be sceptical about subgroup effects, and the burden of proof lies with those claiming such effects. Nevertheless, true subgroup effects do sometimes exist, and Core GRADE users require methods to identify such instances and distinguish them from spurious associations.

Criteria for judging the credibility of subgroup effects

For almost 50 years methodologists and statisticians have been writing about how to distinguish credible from spurious subgroup claims.[20] In the following, we apply the key lessons from this inquiry to an example.

In an exploration of subgroup effects, authors postulated that randomised trials of β blockers showing greater reductions in heart rate would show larger relative risk reductions in deaths among patients with heart failure.[21] The authors found an apparent effect modification: for every five beats per minute reduction in heart rate with β blocker treatment, they found a commensurate 18% reduction in the risk of death. The question arises: is this a true or spurious subgroup effect?

In deciding on the credibility of subgroup effects, one issue specific to systematic reviews and meta-analyses is whether the effect modification was based on a comparison between studies (eg, β blockers achieved different reductions in heart rate in different studies and this is the basis of the analysis) or a within study comparison (the same study included interventions with greater and lesser heart rate reduction, achieved, for example, by including groups with larger and smaller doses of β blockers). Within study comparisons are far more compelling than between study comparisons. In this case, however, the analysis relies exclusively on between study comparisons, reducing the credibility of the apparent effect modification.

Perhaps the most important single issue in addressing a putative subgroup effect is whether chance can explain the difference in effect between subgroups. The lower the P value associated with the appropriate statistical test—referred to as a test of interaction—the less likely chance is an explanation and the more credible becomes the postulated effect.

However, this statistical criterion can be severely undermined if authors have not prespecified subgroup analyses, have conducted a large number of subgroup analyses, or report only selected results. Violation of any of these criteria greatly increases the probability that chance rather than a true subgroup effect is responsible for apparent differences between groups, and thus renders the P value associated with the test of interaction far less trustworthy. In this case the authors specified the subgroup analysis in advance but tested 12 hypotheses with a P value of 0.006 for interaction.

Recently, a team of methodologists developed the first formal Instrument for assessing the Credibility of Effect Modification ANalyses (ICEMAN, www.iceman.help).[22] This instrument addresses all the issues we have discussed, along with several others, and is straightforward to apply. Supplementary appendix 2 presents the full ICEMAN related assessment that led to a conclusion of moderate credibility of the authors' subgroup hypothesis.

Addressing the results of the subgroup credibility exploration

If Core GRADE users conclude that the putative subgroup effect is of low or very low credibility they will present results only for the summary of all studies, rating inconsistency for the entire population. However, a conclusion of moderate or high credibility warrants the creation of separate PICO questions for each subgroup, separate presentation of results for each subgroup, separate ratings of certainty considering all five domains of rating down, and separate conclusions in keeping with each estimate of effect.

A result near the threshold between low and moderate credibility presents challenges. One option is to present both the overall and the subgroup results in the summary of findings table. A second is to present only one of the overall and subgroup results in the summary of findings table and report, in the text, a briefer summary of the one not chosen for the summary of findings table. Whatever they choose, authors should acknowledge the close-call nature of the credibility assessment.

In the example of β blockers to reduce mortality in patients with heart failure, the conclusion regarding credibility falls in the range of moderate credibility. Because the effect modifier was a continuous variable, the authors chose, rather than an arbitrary threshold, the more powerful continuous meta-regression approach to the analysis. Their results thus suggest that the greater the effect in reducing heart rate, the greater the mortality reduction. The moderate credibility of the effect suggests possible results of shared decision making with patients and their clinicians: use doses of β blockers that substantially but safely reduce the patients' heart rate.

Conclusion

When Core GRADE users construct PICO frameworks that are broad with respect to both patients and interventions—as we believe they should—they must prepare for the possibility of inconsistent results. They do so by identifying a priori hypotheses to explain inconsistency, including a postulated direction.

Having decided on their subgroup hypotheses, Core GRADE users address the key criteria for evaluating inconsistency. Examining the forest plot, they note the magnitude of differences in point estimates, the extent to which the CIs overlap, and where the point estimates lie in relation to the target of their certainty rating. The greater the variability in point estimates and the less the overlap of CIs, the more likely there is problematic inconsistency. The decision, however, requires consideration of the chosen threshold for certainty rating: whether the null or the MID, the greater the extent to which, in the presence of minimally overlapping CIs, point estimates fall on opposite sides of the threshold, the more likely there is problematic inconsistency.

Problematic inconsistency requires determining if a priori hypotheses can explain that inconsistency. Critical criteria for judging the credibility of any apparent subgroup effects include whether the analysis is based on within trial or between trial comparisons, the P value of a test of interaction, and whether the analysis is based on a small number of a priori hypotheses with a specified direction. If the subgroup effect proves credible, Core GRADE users will provide separate evidence summaries for each subgroup and rate certainty of evidence accordingly. If not, they will assess inconsistency across all eligible studies.

AUTHOR AFFILIATIONS

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

[2]Department of Medicine, McMaster University, Hamilton, ON, L8S 4L8, Canada

[3]MAGIC Evidence Ecosystem Foundation, Oslo, Norway

[4]Division of Clinical Epidemiology, University Hospital and University of Basel, Basel, Switzerland

[5]School of Public Health, University College Cork, Cork, Ireland

[6]MTA–PTE Lendület "Momentum" Evidence in Medicine Research Group, Medical School, University of Pécs, Pécs, Hungary

[7]Guide2Guidance, Lemelerberg 7, Utrecht, Netherlands

[8]Clinical Research and Epidemiology, Institute of Liver and Biliary Sciences, New Delhi, India

[9]Evidence-based Practice Center, Mayo Clinic, Rochester, MN, USA

[10]Peninsula Dental School, University of Plymouth, Plymouth, UK

[11]Faculty of Dentistry, Manipal University College Malaysia, Melaka, Malaysia

[12]Department of Obstetrics and Gynecology, McMaster University, Hamilton, ON, Canada

[13]Division General Internal Medicine, University Hospitals of Geneva, Geneva, Switzerland

[14]Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore

[15]Department of Medicine, University of Kansas Medical Center, Kansas City, MO, USA

[16]Institute of Health and Society, University of Oslo Faculty of Medicine, Oslo, Norway

[17]HTA Region Stockholm, Centre for Health Economics, Informatics and Health Care Research (CHIS), Stockholm Health Care Services, Stockholm, Sweden

[18]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

[19]Division of Endocrinology, Department of Medicine, Mayo Clinic, Rochester, MN, USA

[20]Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN, USA

Contributors: GG, VMM, TA, MH, and AI conceived and designed the Core GRADE series. GG, SS, VMM, RB-P, MH, HDB, and MP drafted this article. All authors critically revised the article across several iterations for important intellectual content and gave final approval for the article. GG is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: No external funding.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Gyuatt G, Agoritsas T, Brignardello-Petersen R, et al. Core GRADE 1: overview of the Core GRADE approach. *BMJ* 2025;389:e081903.

2 Guyatt G, Zeng L, Brignardello-Petersen R, et al. Core GRADE 2: choosing the target of certainty rating and assessing imprecision. *BMJ* 2025;389:e081904.

3 Hatala R, Keitz S, Wyer P, Guyatt G, Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine: 4. Assessing heterogeneity of primary studies in systematic reviews and whether to combine their results. *CMAJ* 2005;172:661-5. doi:10.1503/cmaj.1031920.

4 Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2002;21:1575-600. doi:10.1002/sim.1188.

5 Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. *Int J Epidemiol* 2002;31:72-6. doi:10.1093/ije/31.1.72.

6 Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med* 1998;17:1923-42. doi:10.1002/(SICI)1097-0258(19980915)17:17‹1923::AID-SIM874›3.0.CO;2-6.

7 Torres Roldan VD, Ponce OJ, Urtecho M, et al. Understanding treatment-subgroup effect in primary and secondary prevention of cardiovascular disease: An exploration using meta-analyses of individual patient data. *J Clin Epidemiol* 2021;139:160-6. doi:10.1016/j.jclinepi.2021.08.006.

8 Hanlon P, Butterly EW, Shah AS, et al. Treatment effect modification due to comorbidity: Individual participant data meta-analyses of 120 randomised controlled trials. *PLoS Med* 2023;20:e1004176. doi:10.1371/journal.pmed.1004176.

9 Rhodes KM, Turner RM, Higgins JP. Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. *Res Synth Methods* 2016;7:346-70. doi:10.1002/jrsm.1193.

10 Agarwal A, Hunt B, Stegemann M, et al. Therapeutics and COVID-19: living guideline 2023. https://app.magicapp.org/#/guideline/6989. [Accessed 20 Sept 2024.]

11 Carson JL, Stanworth SJ, Guyatt G, et al. Red Blood Cell Transfusion: 2023 AABB International Guidelines. *JAMA* 2023;330:1892-902. doi:10.1001/jama.2023.12914.

12 Uthman OA, Okwundu C, Gbenga K, et al. Optimal Timing of Antiretroviral Therapy Initiation for HIV-Infected Adults With Newly Diagnosed Pulmonary Tuberculosis: A Systematic Review and Meta-analysis. *Ann Intern Med* 2015;163:32-9. doi:10.7326/M14-2979.

13 Luetkemeyer AF, Kendall MA, Nyirenda M, et al, Adult AIDS Clinical Trials Group A5221 Study Team. Tuberculosis immune reconstitution inflammatory syndrome in A5221 STRIDE: timing, severity, and implications for HIV-TB programs. *J Acquir Immune Defic Syndr* 2014;65:423-8. doi:10.1097/QAI.0000000000000030.

14 Guyatt G, Zhao Y, Mayer M, et al. GRADE guidance 36: updates to GRADE's approach to addressing inconsistency. *J Clin Epidemiol* 2023;158:70-83. doi:10.1016/j.jclinepi.2023.03.003.

15 Karlsen AP, Wetterslev M, Hansen SE, Hansen MS, Mathiesen O, Dahl JB. Postoperative pain treatment after total knee arthroplasty: A systematic review. *PLoS One* 2017;12:e0173107. doi:10.1371/journal.pone.0173107.

16 Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539-58. doi:10.1002/sim.1186.

17 Borenstein M. In a meta-analysis, the I-squared statistic does not tell us how much the effect size varies. *J Clin Epidemiol* 2022;152:281-4. doi:10.1016/j.jclinepi.2022.10.003.

18 Borenstein M, Higgins JP, Hedges LV, Rothstein HR. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Res Synth Methods* 2017;8:5-18. doi:10.1002/jrsm.1230.

19 Alba AC, Alexander PE, Chang J, MacIsaac J, DeFry S, Guyatt GH. High statistical heterogeneity is more frequent in meta-analysis of continuous than binary outcomes. *J Clin Epidemiol* 2016;70:129-35. doi:10.1016/j.jclinepi.2015.09.005.

20 Schandelmaier S, Chang Y, Devasenapathy N, et al. A systematic survey identified 36 criteria for assessing effect modification claims in randomized trials or meta-analyses. *J Clin Epidemiol* 2019;113:159-67. doi:10.1016/j.jclinepi.2019.05.014.

21 McAlister FA, Wiebe N, Ezekowitz JA, Leung AA, Armstrong PW. Meta-analysis: beta-blocker dose, heart rate reduction, and death in patients with heart failure. *Ann Intern Med* 2009;150:784-94. doi:10.7326/0003-4819-150-11-200906020-00006.

22 Schandelmaier S, Briel M, Varadhan R, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. *CMAJ* 2020;192:E901-6. doi:10.1503/cmaj.200077.

Supplementary information: Appendix 1

Supplementary information: Appendix 2