Core GRADE 3: Rating Certainty of Evidence – Assessing Inconsistency

Gordon Guyatt, distinguished professor [1 2 3], Stefan Schandelmaier, methodologist [4 5 6], Romina Brignardello-Petersen, associate professor [1], Hans De Beer, methodologist [7], Manya Prasad, associate professor [8], M Hassan Murad, professor [9], Prashanti Eachempati, adjunct Professor [3 10 11], Derek K. Chu, assistant professor [1 2], Rohan D'Souza, associate professor [1 12], Alfonso Iorio, professor [1 2], Thomas Agoritsas, asscociate professor [1 3 13], Liang Yao, assistant professor [14], Reem A Mustafa, professor [1 15], Sameer Parpia, associate professor [1], Pasqualina Santaguida, assistant professor [1], Per Olav Vandvik, professor [3 16 17], Monica Hultcrantz, head of HTA Region Stockholm [18 19], Victor Montori, professor [20 21]

1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
2. Department of Medicine, McMaster University, Hamilton, Ontario, Canada
3. MAGIC Evidence Ecosystem Foundation, Oslo, Norway
4. 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, 3524 LC Utrecht, the Netherlands
8. Clinical Research and Epidemiology, Institute of Liver and Biliary Sciences, New Delhi, India
9. Evidence-based Practice Center, Mayo Clinic, Rochester, MN 55905, USA
10. Peninsula Dental School, University of Plymouth, United Kingdom
11. Faculty of Dentistry, Manipal University College Malaysia, Melaka, Malaysia
12. Department of Obstetrics & Gynecology, McMaster University, Hamilton, Ontario, 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. Department of Medicine, Lovisenberg Diakonale Hospital, Oslo, Norway
18. HTA Region Stockholm, Centre for Health Economics, Informatics and Health Care Research (CHIS), Stockholm Health Care Services, Sweden
19. Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden
20. Division of Endocrinology, Department of Medicine, Mayo Clinic, Rochester, MN, USA
21. Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN, USA

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

Abstract

This third article in a seven-part series presents the Core GRADE approach to deciding whether to rate down the certainty of evidence due to inconsistency – that is, unexplained variability in results across studies. For continuous outcomes, in which we generally have only absolute measures of effect – typically mean differences - we assess consistency in those absolute effects. For binary outcomes, because relative effects such as risk ratios are usually constant across different levels of baseline risk while absolute effects vary, we typically consider consistency only with respect to relative and not absolute effects.

First, planning for the possibility of inconsistent results across studies, systematic review authors using GRADE will construct a priori hypotheses regarding possible subgroup effects - synonym, effect modification - including population or intervention characteristics that may explain inconsistency. Second, having made these hypotheses, they will review results and judge the extent of inconsistency. This judgement involves a careful visual assessment of the relevant forest plot.

In general, the greater the extent to which the point estimates differ, and the less the overlap in confidence intervals, the more compelling the grounds to rate down certainty of evidence. However, 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): even when point estimates vary considerably, if most lie on the same side of the chosen threshold, they are unlikely to rate down for inconsistency.

Finally, review authors will test their subgroup hypothesis using established criteria to determine the credibility of any apparent subgroup effects. If an effect proves credible, they will provide separate evidence summaries and rate certainty of evidence separately for each subgroup. When, on the other hand, authors find no credible subgroup effect they will provide a single evidence summary, rating down for inconsistency if necessary. Recommendations would then apply to the entire population.

1. Introduction

This is the third in a series of papers describing the essentials of GRADE in rating certainty of evidence and grading recommendations for paired interventions and comparators focusing on the perspective of the patient and clinician. *The prior two papers provided an overview of the 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 differences in results between studies. We are particularly concerned about inconsistency 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. Readers may find useful an elaboration of issues of heterogeneity in a prior paper. [3]

To best address inconsistency, Core GRADE users must first understand the measure of effect on which they should focus. Users should focus on relative effects such as risk ratios or hazard ratios when dealing with binary outcomes and absolute effects such as mean differences when dealing with continuous outcomes. Next, they must prepare for the possibility that they will encounter large inconsistency by making a priori hypotheses that may 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 existing inconsistency. If after considering these hypotheses, large unexplained inconsistency remains, they will rate down the certainty of the evidence. The paper takes readers through these steps.

*** Choose a priori hypotheses with a specified direction, for binary outcomes hypotheses deal with relative effects.

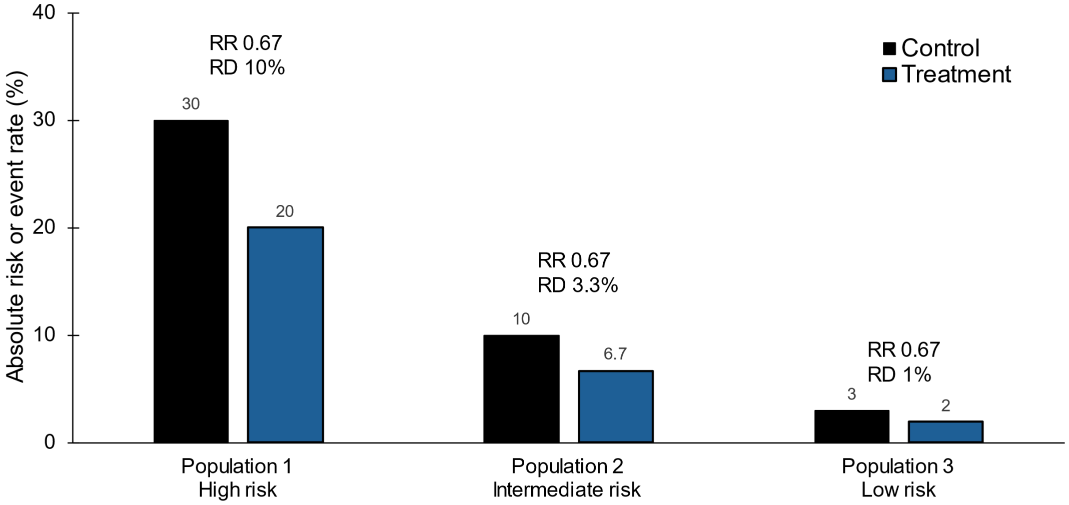
2. Choosing the right measure of effect when assessing inconsistency

2.1 Binary outcomes: variability in relative versus absolute effects

As we 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 and more sick.[4-9] However, given that such patient characteristics are often associated with substantial differences in baseline risk (i.e. 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 – 33% – across low, medium and high-risk groups. Because of the very 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.

Figure 1. Constant relative risk with varying baseline risk, leading to varying reduction in absolute risks. RR stands for relative risk and RD for risk difference. In each population, the larger event rate represents the control (baseline risk, black bars) and the smaller event rate the intervention (blue bars).



Despite risk differences being more important to patients than relative risks, authors of randomized 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 highlighted in Figure 1. Greater consistency in results is desirable: it increases our confidence in the pooled estimates of effect. Thus, the anticipated consistency of relative and variability of absolute effects is the reason why in GRADE Summary of Findings tables we estimate risk differences in each relevant patient group by applying relative risks to baseline risks, and why we may provide different treatment recommendations for low-, medium- and high-risk individuals. 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.

*** For binary outcomes, specify risk groups for establishing the baseline risk and source of baseline risk information

2.2 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.

3. GRADE's approach to preparing for inconsistency

In this section, we provide a discussion regarding how Core GRADE users, when thinking ahead to possible inconsistency in results, formulate a plan to best deal with the inconsistency they 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 a number of reasons for inconsistency. These include random error and differences in population – intervention – comparator – outcome (PICO) elements. Hypotheses may be able to explain these differences – this is the hope for when we prepare 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 inconsistency judgments within each subgroup. If they do not, the unexplained variability in effects decreases our certainty of evidence.

3.1 Variability in PICO elements

GRADE ratings of certainty pertain to bodies of evidence summarized in rigorous systematic reviews. The GRADE process begins with construction of a structured clinical question.[1] Studies addressing a particular question are certain to vary in patients they enrol, aspects of the intervention and comparator they have chosen, and the way their measurement of the outcome, and this variability is often appreciable.

Core GRADE users may intuit that such variability (i.e., inconsistency in PICO elements), compromises the certainty of evidence from a systematic review. This, however, is very 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.

3.2 Considering possible subgroups with different intervention effects: Three possible options

When reflecting on the possibility that effects differ across patient subgroups (e.g., effects may differ in the old versus young) or across interventions subgroups (e.g., 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 generalizability and precision but runs the risk of pooling inappropriately across patient or interventions subgroups if effects differ substantially.

To solve the problem, Core GRADE users must distinguish, for each subgroup consideration, between three scenarios: 1) one has no reason to suspect differences in effects across subgroups, 2) one is confident that effects vary across subgroups, or 3) one has good reason to suspect subgroup differences but is uncertain.

Think, for example, of patients with different ages. The following are the three scenarios and the corresponding actions they would mandate:

1. Previous research provides little support for the possibility that effects differ in the old and young. 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 highly confident that the relative effects on older vs younger individuals differ. Accordingly, one would choose a narrow age range for the PICO (e.g., older individuals) or create two separate PICOs and sets of recommendations, one for older individuals and the other for younger individuals.
3. Previous research plausibly suggests that effects differ in the old and the young 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 summarizes the three scenarios when considering subgroups during PICO construction and provides examples of each.

<u>Table 1.</u> Three scenarios when considering subgroups during PICO construction.

Scenario	Implications for PICO construction	Example
1. Previous research provides no compelling evidence that effects should be different across patient or intervention subgroups (no subgroup hypothesis)	Combine all subgroups (single estimate of effect) without a subgroup hypothesis	The World Health Organization (WHO) has generated a number of recommendations regarding management of COVID-19 patients. 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 suggest 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 anemic patients considered that the biology differed in 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 concomitant diagnosis of HIV and tuberculosis tested whether the impact of early versus delayed treatment on mortality differed in those with higher and lower CD4+ cell counts.[12] A previous trial suggested that hypothesis including a clear direction, but for another outome.[13]

We recommend that review authors, to maximize precision and generalizability, 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, when constructing the PICO, choosing the third scenario. We now present details of how to deal with this third scenario.

3.3 Need for a priori hypotheses with a specified direction

As we pointed out in the first article in this series, 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 (e.g. older versus younger, more- versus less sick) or differences in interventions (e.g. 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 prior evidence (e.g., from a related trial, meta-analysis, or cohort study) or a thorough understanding of the underlying biology and include the direction of the subgroup effect (hypothesising, for example, not just that effects may differ across patient age, but also that effects will be larger in the old than in the young). Postulating more than a small number (ideally three or fewer) of directional hypotheses will increase the likelihood of chance findings (spurious associations), thus undermining the credibility of any subgroup effects.

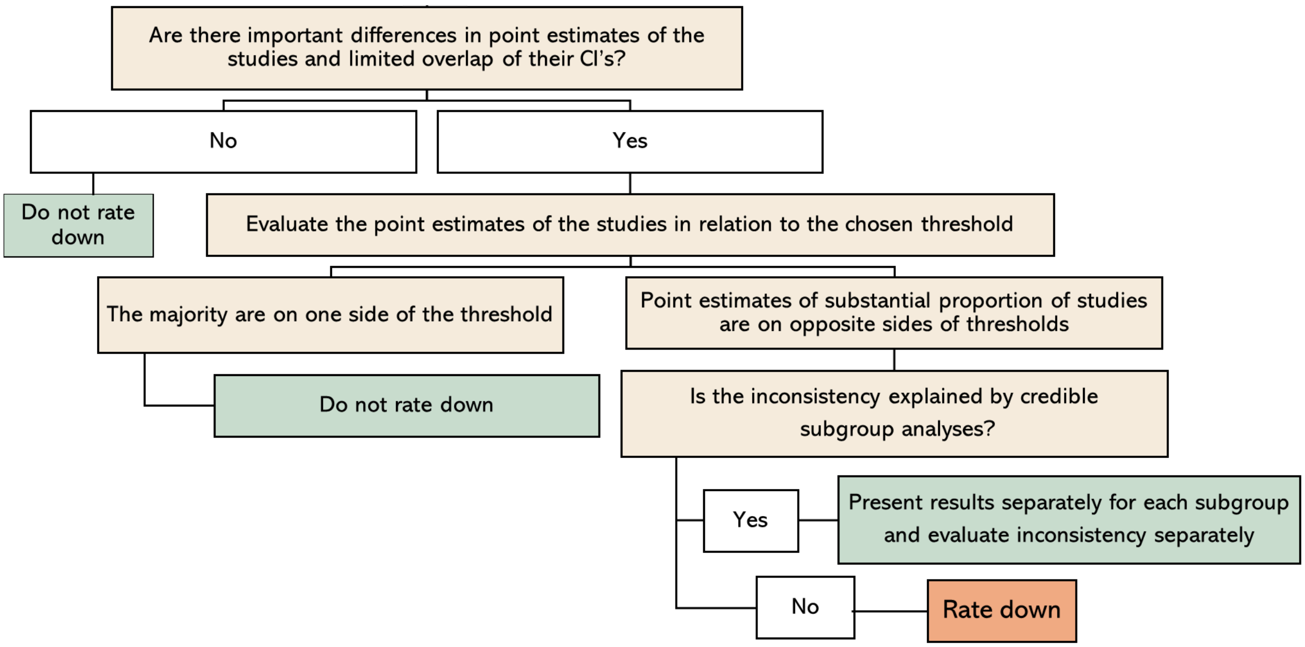
For instance, in the systematic review of when to start anti-retroviral therapy in patients with concomitant diagnoses of tuberculosis and HIV, in considering mortality, the authors made only a single a priori hypothesis. They postulated that effects may differ depending on CD4$^+$ T cell counts using a threshold of <0.050 x 10^9 cells/L vs. >0.050 x 10^9 cells/L.[11 12] Their hypothesis was based on prior 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, the results suggested, contrary to the hypothesis, that if there was a benefit of early therapy it was more likely in those with a low cell count (interaction p = 0.12). The example thus highlights how the review authors prepared themselves for the the possibility of inconsistent results through specifying a single, directed subgroup hypothesis based on related evidence. The example also highlights that caution is advisable before concluding, without a subgroup analysis, that effects differ between subgroups.

The ability to predict the direction of a subgroup effect provides a useful criterion when deciding between option 1 (broad PICO, no subgroup analysis) and option 3 (broad PICO and subgroup analysis) in the previous section. If one cannot confidently specify the direction of the potential subgroup effect, one should choose option 1 rather than option 3. Consistent with our recommendation of a small number of compelling subgroup hypotheses, we discourage post hoc exploration of possible subgroup effects.

4. Criteria for judging serious inconsistency

Having addressed how Core GRADE users should plan for dealing with the inconsistency in results they find, we will now address how they will implement their plan (see Figure 2). In sections 4.1 to 4.3 we describe how Core GRADE users can determine whether there is sufficient inconsistency to be concerned and 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. Section 5 deals with this issue of subgroup explanations of variability in results. If there is only one eligible study, Core GRADE users will not rate down for inconsistency, though if the authors provide the data, they may still address the possibility of subgroup effects.

<u>Figure 2</u>: A flow chart summarizing Core GRADE's approach to addressing inconsistency in results


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

4.1 Three visual criteria from forest plots

Consider the hypothetical body of evidence in Figures 2A and 2B. When considering whether studies yield similar or different results, most observers of these forest plots will quickly conclude that results from 2A are very consistent while the results from 2B are inconsistent. What aspects of the results justify these inferences?

i) Point estimates:

One is more inclined to consider rating down for inconsistency where point estimates differ substantially between studies.

In 2A, the point estimates are very similar, ranging from 0.71 to 0.76. The similarity in the point estimates suggests there is no need to consider rating down for inconsistency.

In 2B, in contrast, two studies suggest substantial treatment effects - relative risk reductions of over 50% - and two others suggest modest harms, 17% and 25% increases in relative risk. The large differences in the point estimates of the two pairs of studies suggests rating down for inconsistency.

ii) Overlap of confidence intervals:
One is more inclined to consider rating down for inconsistency if the confidence intervals of included studies do not demonstrate substantial overlap.

In 2A, the confidence intervals of the four studies are largely overlapping. This overlap suggests there is no need to consider rating down for inconsistency.

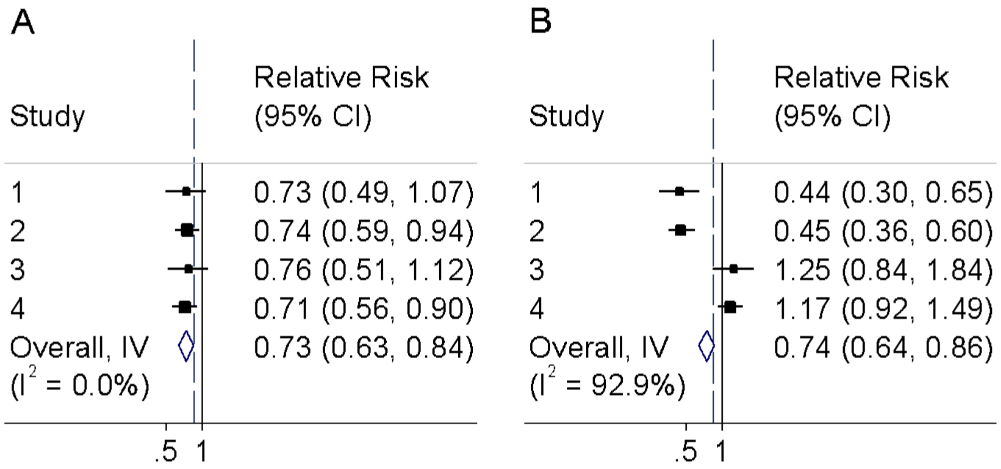
In 2B, in contrast, the confidence intervals between the first and second pairs of studies are completely non-overlapping. This provides a strong rationale for rating down for inconsistency.

iii) Relation of point estimates to 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 (i.e. the null – that is, no difference between intervention and comparator - or minimal important difference [MID] [2]). In these situations, they will be less inclined to rate down for inconsistency.

Whichever threshold one uses, in Figure 2A all studies are on one side of the threshold (no need to consider rating down for inconsistency). In Figure 2B, the pairs of studies are on opposite sides of either threshold with one pair demonstrating benefit and the other demonstrating harm (thus the need to consider rating down) and, crucially, with no overlap in the confidence intervals.

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 approximately 15% will translate into a minimally important absolute effect of 1%. Appendix 1 describes this process.

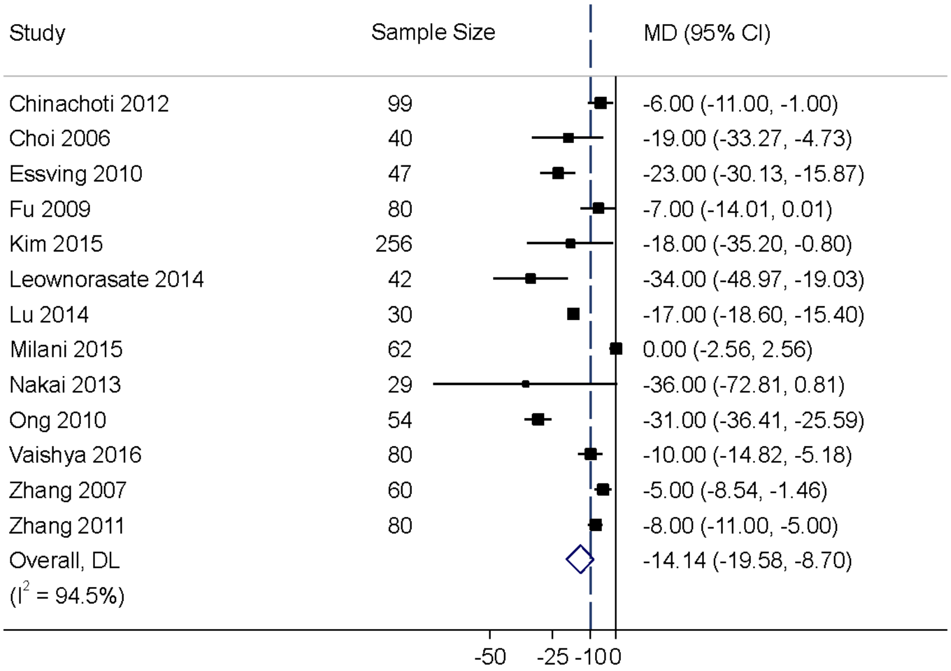
Figure 2. Forest plots of (A) consistent, and (B) inconsistent results from four randomized trials with similar overall pooled effects. The broken line represents the minimal important difference.



4.2 Applying visual criteria: an example highlighting how the choice of thresholds affects judgments of inconsistency

The three key criteria for judging inconsistency – similarity of point estimates, overlap of confidence intervals and relation of results to our chosen threshold for rating certainty – apply equally well to continuous outcomes. Consider Figure 3 that depicts 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 issues.[14 15]

__Figure 3__. Impact of local infiltration analgesia on chronic pain: Forest plot from a systematic review. The broken vertical line represents an estimate of the MID in pain score (10 mm) on a 100 mm visual analogue scale.



VAS scale range 0-100, with higher denoting more pain. Long-dash blue line represents MID of 10 points.

Consider the appropriate inference if authors of the systematic review of this evidence choose 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.

What if, however, review authors chose to rate their certainty with respect to the MID and chose 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.

While 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 there is non-overlap in confidence intervals, they will seldom find compelling reason to invoke this additional criterion.

4.3 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 Figure 2A. 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 Figure 2B in which the I^2 is 93%.

Unfortunately, I^2 may prove misleading. [17-19] In particular, if the included studies have narrow confidence intervals 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 3, the extremely 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 favoring the intervention demonstrating 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 under 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 a theoretical possibility, we have found instances of compelling reason to rate down twice for inconsistency sufficiently unusual that it need not concern users of Core GRADE.

5. Examining the credibility of apparent subgroup effects based on a priori hypotheses

5.1 The burden of proof lies on those claiming a subgroup effect

We have pointed out that, overwhelmingly, relative effects 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 skeptical 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.

5.2 Criteria for judging the credibility of subgroup effects

Methodologists and statisticians have been writing for almost 50 years 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 randomized trials of beta-blockers resulting in greater reductions in heart rate would demonstrate 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 beta-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 between-study comparison (e.g. 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 the postulated effect becomes.

This statistical criterion can, however, be severely undermined if authors have not specified subgroup analyses beforehand, have conducted a large number of subgroup effects, or report only selected results. Violation of any of these criteria greatly increase the probability that chance, rather than a true subgroup effect, is responsible for apparent difference between groups. They therefore render 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: the p-value for interaction proved 0.006.

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

5.3 What to do with the results of the subgroup credibility exploration?

If one concludes that the putative subgroup effect is of low or very low credibility one dismisses further consideration and presents results only for the summary of all studies, rating inconsistency for the entire population. On the other hand, 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 (SoF) table. A second is to present only one of the overall and subgroup results in the SoF and report a briefer summary of the one not chosen for the SoF in the text. 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, rather than choosing an arbitrary threshold, the authors chose 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 provides evidence that clinicians can suggest to their patients using doses of β -blockers that substantially but safely reduce the patients' heart rate.

6. 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 confidence intervals 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 confidence intervals, the more likely problematic inconsistency exists. 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 confidence intervals, point estimates fall on opposite sides of the threshold, the more likely problematic inconsistency exists.

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- or between-trial comparisons, the p-value 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.

References

1. Gyuatt G, Agoritsas T, Brignardello-Petersen R, et al. Core GRADE 1: Overview of the Core GRADE Process *BMJ (in submission)* 2024
2. Guyatt G, Zeng L, Brignardello-Petersen R, et al. Core GRADE 2: Choosing the Target of Certainty Rating and Assessing Imprecision. *BMJ (in submission)* 2024
3. Hatala R, Keitz S, Wyer P, et al. 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(5):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(11):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(1):72-6. doi: 10.1093/ije/31.1.72
6. Schmid CH, Lau J, McIntosh MW, et al. 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(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;139:160-66. doi: 10.1016/j.jclinepi.2021.08.006 [published Online First: 20210813]
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(6):e1004176. doi: 10.1371/journal.pmed.1004176 [published Online First: 20230606]
9. Rhodes KM, Turner RM, Higgins JP. Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. *Res Synth Methods* 2016;7(4):346-70. doi: 10.1002/jrsm.1193 [published Online First: 20151217]
10. Agarwal A, Hunt B, Stegemann M, et al. Therapeutics and COVID-19: living guideline 2023 [Available from: https://app.magicapp.org/#/guideline/6989 accessed Sept 20 2024.
11. Carson JL, Stanworth SJ, Guyatt G, et al. Red Blood Cell Transfusion: 2023 AABB International Guidelines. *Jama* 2023;330(19):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(1):32-9. doi: 10.7326/m14-2979
13. Luetkemeyer AF, Kendall MA, Nyirenda M, et al. Tuberculosis immune reconstitution inflammatory syndrome in A5221 STRIDE: timing, severity, and implications for HIV-TB programs. *J Acquir Immune Defic Syndr* 2014;65(4):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 [published Online First: 20230309]
15. Karlsen AP, Wetterslev M, Hansen SE, et al. Postoperative pain treatment after total knee arthroplasty: A systematic review. *PLoS One* 2017;12(3):e0173107. doi: 10.1371/journal.pone.0173107 [published Online First: 20170308]
16. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21(11):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;152:281-84. doi: 10.1016/j.jclinepi.2022.10.003 [published Online First: 20221009]
18. Borenstein M, Higgins JP, Hedges LV, et al. Basics of meta-analysis: I(2) is not an absolute measure of heterogeneity. *Res Synth Methods* 2017;8(1):5-18. doi: 10.1002/jrsm.1230 [published Online First: 20170106]

19. Alba AC, Alexander PE, Chang J, et al. 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 [published Online First: 20150918]
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 [published Online First: 20190524]
21. McAlister FA, Wiebe N, Ezekowitz JA, et al. Meta-analysis: beta-blocker dose, heart rate reduction, and death in patients with heart failure. *Ann Intern Med* 2009;150(11):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(32):E901-e06. doi: 10.1503/cmaj.200077

Contributors: Gordon Guyatt, Victor Montori, Thomas Agoritsas, Monica Hultcrantz and Alfonso Iorio conceived and designed the Core GRADE series. Gordon Guyatt, Stefan Schandelmaier, Victor Montori, Romina Brignardello-Petersen, Monica Hultcrantz, Hans DeBeer, and Manya Prasad drafted this article. All authors critically revised the article across several iterations for important intellectual content and gave final approval for the article. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Competing interests: All authors have completed the ICMJE uniform disclosure form and declare: no financial support from any industry for the submitted work; no other relationships or activities that could appear to have influenced the submitted work.

Funding: The Core GRADE series did not receive any external funding.

Patient and public involvement: We did not include patients and public.