Core GRADE 4: Rating certainty of evidence — risk of bias, publication bias, and reasons for rating up certainty

Gordon Guyatt, distinguished professor,[1 2 3 &] Ying Wang, methodologist,[1 &] Prashanti Eachempati, adjunct Professor,[3 4 5] Alfonso Iorio, professor,[1 2] M Hassan Murad, professor,[6] Monica Hultcrantz, head of HTA Region Stockholm,[7 8] Derek K. Chu, assistant professor, [1 2] Ivan D. Florez, Professor,[9 10 11] Lars G. Hemkens, professor,[12 13 14] Thomas Agoritsas, associate professor,[1 3 15] Liang Yao, assistant professor,[16] Per Olav Vandvik, professor,[3 17 18] Victor Montori, professor,[19 20] Romina Brignardello-Petersen, associate professor [1]

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. Peninsula Dental School, University of Plymouth, United Kingdom
5. Faculty of Dentistry, Manipal University College Malaysia
6. Evidence-based Practice Center, Mayo Clinic, Rochester, MN 55905, USA
7. HTA Region Stockholm, Centre for Health Economics, Informatics and Health Care Research (CHIS), Stockholm Health Care Services, Sweden
8. Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden
9. Department of Pediatrics, University of Antioquia, Medellin, Colombia
10. Pediatric Intensive Care Unit, Clínica Las Américas-AUNA, Medellin, Colombia
11. School of Rehabilitation Science, McMaster University, Hamilton, Canada
12. Pragmatic Evidence Lab, Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland
13. Department of Clinical Research, University Hospital Basel and University of Basel, Basel, Switzerland
14. Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, USA
15. Division General Internal Medicine, University Hospitals of Geneva, Geneva, Switzerland
16. Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore
17. Institute of Health and Society, University of Oslo Faculty of Medicine, Oslo, Norway
18. Department of Medicine, Lovisenberg Diakonale Hospital, Oslo, Norway
19. Division of Endocrinology, Department of Medicine, Mayo Clinic, Rochester, MN, USA
20. Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN, USA

Gordon Guyatt and Ying Wang are the co-first authors.

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

In Core GRADE, randomized trials begin as high certainty evidence and non-randomized studies of interventions (NRSI) begin as low certainty evidence. To decide whether to rate down certainty of evidence for risk of bias, Core GRADE users conducting systematic reviews begin by assessing risk of bias for each individual study. Key risk of bias issues for randomized trials include allocation concealment, blinding, and loss to follow-up. Key issues for NRSI include accurate measurement of confounders and an appropriately adjusted analysis. Using pre-determined criteria, Core GRADE users will classify individual studies as at overall low or high risk of bias.

Systematic review authors will then determine whether results of eligible studies are similar in the studies at low and high risk of bias. If they differ substantially, Core GRADE users will use only the results from low risk of bias studies as the best estimates of effect and not rate down certainty of the evidence. If results between low and high risk of bias studies are not substantially different, they will use the pooled results from all studies to inform their best estimate. In such instances, if high risk of bias studies dominate the body of evidence, Core GRADE users will rate down the certainty of the entire body of evidence for risk of bias. Otherwise, they will not rate down.

With respect to publication bias, a body of evidence comprising relatively small studies most or all funded by industry providers should raise suspicion of publication bias. Other approaches include visual inspection of funnel plot asymmetry and statistical tests. Because of the limitations of the approaches for assessing publication bias, Core GRADE users will often be left with uncertainty. Core GRADE therefore suggests using the terms “undetected” or “strongly suspected” to describe publication bias.

Core GRADE users appraising results from well conducted NSRI can consider rating up the certainty of evidence for one level when pooled estimates demonstrate risk ratios greater than 2.0 or less than 0.5. When the risk ratio is greater than 5.0 or less than 0.2 (very large effect), Core GRADE users can consider rating up two levels. Similarly, when a credible dose-response gradient exists, they can also consider rating up certainty of evidence for NRSI.

Curricular Objectives

After reading this paper, Core GRADE users will be able to:

- Understand the definition of risk of bias.
- Choose appropriate instruments for assessing risk of bias of individual studies.
- Rate risk of bias across the body of evidence by considering the similarities or differences in results from low and high risk of bias studies as well as their relative contribution to the pooled estimate.
- Understand the causes of and approaches for detecting publication bias.
- Make appropriate judgements regarding when to rate up the certainty of evidence based on non-randomized studies of interventions (NRSI).

Take Home Messages

- To decide whether to rate down certainty of evidence for risk of bias, Core GRADE users begin by applying criteria that facilitate classifying individual studies as overall low or high risk of bias.
- If low and high risk of bias studies suggest substantially different effects, Core GRADE users will in general use only the results from low risk of bias studies and not rate down certainty of evidence for risk of bias.
- If results do not differ substantially between low and high risk of bias studies, Core GRADE users will use the pooled results from all studies, and rate down certainty for risk of bias if high risk of bias studies dominate the evidence; otherwise, they will not rate down.
- Considering funnel plot asymmetry or statistical approaches and the role of the pharmaceutical industry in generating the evidence, Core GRADE users can assess risk of publication bias as “undetected” (not rate down for publication bias) or “strongly suspected” (rate down).
- When methodologically rigorous non-randomized studies of interventions (NRSI) suggest a large magnitude of effect or a credible dose-response gradient, Core GRADE users can consider rating up certainty of evidence.

1. Introduction

This is the fourth in a series of papers introducing Core GRADE, the essentials of the GRADE approach to rating certainty of evidence and grading strength of recommendations in systematic reviews, clinical practice guidelines and health technology assessments (HTA). The prior articles presented an overview of the GRADE process and discussed specific aspects of imprecision and inconsistency.

In GRADE’s four-category system of high, moderate, low and very low certainty evidence, randomized controlled trials (RCTs) start as high certainty and non-randomized studies of interventions (NRSI, synonymous with observational studies) start as low certainty. This paper deals with how GRADE addresses reasons for rating down the certainty of evidence for risk of bias, publication bias, and reasons for rating up certainty in NRSI.

2. Risk of Bias

We define bias as a systematic deviation from the underlying true effect of an intervention on an outcome of interest in a given population. Both RCTs and NRSI may be subject to design or execution limitations that may bias the results. Well-designed studies will institute safeguards that minimize risk of bias such as centralized randomisation and blinding. To the extent studies do not implement these safeguards, risk of bias increases. If serious limitations exist among the studies dominating the pooled estimate of effect, Core GRADE users will rate down the overall certainty of evidence for risk of bias.

Issues of risk of bias, and thus safeguards against risk of bias, differ in RCTs and NRSI. We will first deal with RCTs and then with NRSI. The subsequent discussion will address how Core GRADE users should look across the body of evidence to decide whether or not to rate down for risk of bias.

3. Risk of Bias in Individual Studies

3.1 Randomized Trials

Although many instruments exist for addressing risk of bias in parallel group RCTs,[1] a smaller number were designed for general use across all RCTs rather than being tailored to a specific clinical area. These include the Cochrane’s original instrument,[2] their revised instrument (RoB 2),[3] the instrument developed by the Clinical Advances Through Research and Information Translation (CLARITY) group,[4] the Critical Appraisal Skills

Programme (CASP) checklist,[5] the Joanna Briggs Institute (JBI) checklist,[6] the National Institute for Health and Care Excellence (NICE) checklist,[7] the Scottish Intercollegiate Guidelines Network (SIGN) checklist,[8] and an instrument recently developed by an international collaboration of methodologists (named ROBUST-RCT).[9] While each of these instruments possess various limitations, the most important one is that some instruments include items that in Core GRADE are classified as indirectness and imprecision rather than risk of bias, possibly leading to double counting their impact on the GRADE assessment of the certainty of evidence.[1]

Table 1 summarizes the risk of bias items that RCT risk of bias instruments appropriately identify and that Core GRADE users may want to consider.

Table 1. Risk of bias in randomized trials

The most commonly included and important items across various RCT risk of bias tools	● Inadequate generation of random allocation sequence
	● Inadequate concealment of allocation
	● Not blinding participants
	● Not blinding healthcare providers
	● Not blinding data collectors
	● Not blinding outcome assessors
	● Not blinding data analysts
	● Missing outcome data
Other less important items that are variably captured across the RCT risk of bias tools	● Imbalance in co-interventions between groups
	● Difference in outcome assessment or data collection between groups
	● Difference in follow-up time, frequency, or intensity of outcome assessment between groups
	● Deviation from intention-to-treat analysis
	● Selective outcome reporting
	● Early termination for benefit

Two rigorously developed instruments that address limitations of their predecessors' merit particular consideration. One, Cochrane RoB 2,[3] offers a clear process for assessing risk of bias in which signaling questions and algorithms reflect the mechanisms by which bias arises.[3] Cochrane RoB 2 includes five domains addressing randomization, deviations from intended interventions, missing outcome data, outcome measurement, and selective outcome reporting. Cochrane RoB 2 is the only formal instrument that explicitly address cluster and cross-over RCTs.

This instrument has, however, limitations of complexity and difficulty in application.[10] Its sophisticated algorithms and the new terminologies it introduced may contribute to these limitations.[10 11] Studies have reported the low inter-rater reliability of Cochrane RoB 2 and the challenges in implementation that systematic reviewers sometimes experience when using it.[12 13]

A recently developed instrument, the Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT),[9] was inspired by the same motivation as this Core GRADE series: to achieve maximal simplicity without sacrificing important methodological rigor. Strengths of the new instrument include preparatory systematic surveys of existing instruments[1] and of meta-epidemiological studies of risk of bias[14] and extensive pre-testing with both junior and experienced systematic reviewers.

ROBUST-RCT includes six core items addressing random sequence generation, allocation concealment, blinding of participants, blinding of healthcare providers, blinding of outcome assessors, and missing outcome data as well as eight optional items. The instrument provides two approaches to addressing missing outcome data. The more sophisticated approach involves looking across results from all studies. This approach is beyond Core GRADE, but we summarize in Appendix 1 for possible consideration.

Although the items included in ROBUST-RCT have been widely used in other instruments, experience with the ROBUST-RCT presentation remains limited. The instrument is available from the authors and at the time of writing is undergoing peer review.[9] ROBUST-RCT developers provide an Excel file that facilitates its completion. https://www.dropbox.com/scl/fi/gztl83n7fdl6imx1sfmz0/ROBUST-RCT-Excel.xlsx?rlkey=s2nhr203zi3n78ocmz9mi37te&st=7j0rhz9g&dl=0

Both Cochrane RoB 2 and ROBUST-RCT instruments suggest that, in some cases, failure to ensure methodological safeguards may not lead to risk of bias (e.g., blinding of participants is irrelevant in a trial enrolling neonates). Cochrane RoB 2 addresses this issue through the signaling questions in algorithms. ROBUST-RCT does so by including two steps for assessing risk of bias: first, evaluating what happened or whether a methodological safeguard has been implemented (e.g. were participants blinded) and second, judging risk of bias (e.g. did a lack of blinding actually increase bias).

Core GRADE users who particularly appreciate the methodologic sophistication of Cochrane RoB 2 will want to use this instrument for their reviews. Those who put a premium on simplicity and ease of use are likely to choose ROBUST-RCT. There may be Core GRADE users with prior positive experience with one of the other instruments who may value familiarity and continue with its use.

Based on the instrument they choose, Core GRADE users will assess the extent of risk of bias associated with each item for each individual study and subsequently rate each study as low or high risk of bias (see Section 3.4).

3.2 Risk of Bias in Non-Randomized Studies

For assessing certainty of evidence, NRSI as a class of study designs begin as low certainty evidence. Failure to include design features within NRSI that minimize bias can result in further rating down to very low certainty of evidence.

3.2.1 Cohort and Case-Control Studies

When, for a particular outcome, randomized trials do not exist or yield only low or very low certainty evidence, Core GRADE users consider using NRSI for assessing the effects of interventions. NRSI include many study designs, of which the most common are cohort and case-control designs. In cohort studies, investigators compare individuals who have and who have not received a treatment and follow them for the development of the outcomes of interest.[15 16] Case-control studies identify individuals who have and who have not experienced an outcome, and then ascertain whether or not they have received the intervention of interest.[16 17] Table 2 presents key risk of bias issues in NRSI.

Table 2. Risk of bias in non-randomized studies of interventions

● Different eligibility criteria or selection of participants between comparison groups (e.g., participants in intervention and comparator were drawn from different populations)
● Inaccurate measurement of interventions
● Inappropriate measurement of outcome
● Inadequate control of confounders (prognostic factors for outcomes of interest differentially distributed in intervention and control groups) - Inaccurate measurement of confounders - Inadequate adjustment for confounding
● Missing outcome data
● Selective outcome reporting

A large number of instruments are available for assessing risk of bias in NRSI.[18-20] Core GRADE users might consider the relatively simple, straightforward and parsimonious Newcastle-Ottawa quality assessment scale [21] or modifications of that instrument for both cohort and case-control studies developed by the CLARITY group.[22 23]

ROBINS-I ("Risk Of Bias In Non-randomised Studies - of Interventions") version 1 [24] and the revised version 2 [25] represent another option for risk of bias assessments in NRSI. Using the revised version of ROBINS-I, reviewers begin by identifying and listing confounding domains relevant to their study question and then decide whether to proceed with a risk of bias assessment by answering signalling questions aimed at identifying studies at critical risk of bias that would not warrant further assessment. Reviewers then assess bias in seven domains addressing from 3 to 11 signalling questions each. Algorithms based on item responses lead to an overall rating for each domain as either low, moderate, serious or critical risk of bias. Although when using ROBINS-I users begin with NRSI as high certainty evidence, the final ratings with the ROBINS-I and conventional GRADE approaches that start with NRSI at low certainty should align.[26]

Studies have documented that teams often do not use ROBINS-I version 1 correctly,[27] problematic time taken to complete the instrument and its poor usability, misunderstanding of the questions, poor clarity of instructions and overall demanding application.[27-30] These limitations and the instrument's complexity make ROBINS-I an unsatisfactory instrument for Core GRADE, though those who wish to go beyond Core GRADE may still consider its use.

3.2.2 Case Series and Single-Arm Trials

Case series or single-arm trials that include only individuals exposed to an intervention and not those unexposed represent another type of non-randomized study design. Although a tool for assessing risk of bias of case series exists,[31] such assessment is generally not needed when Core GRADE users assess effects of interventions.[32] This is because unbiased assessment of intervention effects requires contemporaneous comparisons of treated to untreated individuals that are lacking in case series. Noncomparative effect estimates are therefore almost always at high risk of bias. Results from single-arm trials are often compared to external controls, typically historical (e.g. comparing survival rates for a new cancer treatment with the survival reported previously with other treatments). Such comparisons are analogous to cohort study designs, but do not allow adjusted analysis, and are thus always at high risk of bias.

Interventions for which harmful effects are restricted to those who receive treatment represent a special case. For instance, only patients who undergo surgery can experience surgical complications. This is also true for other invasive procedures. For example, a study using a large administrative database including over 97,000 individuals who underwent an outpatient colonoscopy identified all those who were admitted to hospital with intestinal bleeding or perforation within 30 days. Because the spontaneous occurrence of such events in any given 24-hour period in individuals not undergoing colonoscopy is almost zero, the study provides an accurate estimate of major complications. Moreover, this study avoided risk of bias issues by accurately documenting adverse events, and thus provides, for colonoscopy adverse events of bleeding and perforation, the same low risk of bias estimates as we find in rigorous RCTs.[33]

3.3 Risk of Bias May Differ Across Outcomes in a Study

Different outcomes from the same study may be at different risk of bias. For instance, risk of bias due to missing outcome data may be higher for an outcome that is difficult to follow for a long period of time (e.g. quality of life) than for an outcome that is easy to follow (e.g. survival). The same is true for risk of bias due to not blinding of outcome assessors: the impact of not blinding on risk of bias may differ across subjective (higher risk) vs objective (lower risk) outcomes.[14] Blinding of outcome assessors is particularly irrelevant for all-cause mortality: for instance, authors of one systematic review noted "Most of the included trials did not blind the outcome assessors; however, mortality can be ascertained without risk of bias".[34]

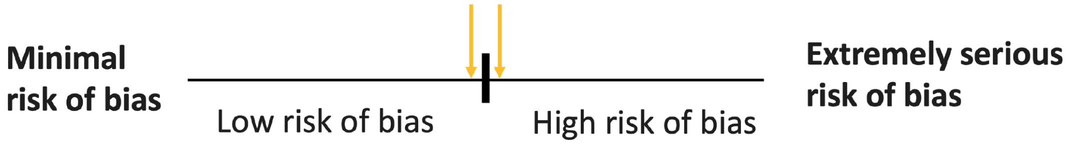
A systematic review evaluating the effect of red and processed meat intake on cardiometabolic and cancer outcomes provides an example of different risk of bias judgements across outcomes in a single study.[35] Considering that risk of bias may differ for all-cause mortality, cardiometabolic outcomes and cancer, authors assessed risk of bias separately for each outcome for each included cohort study. Because prognostic factors differed across outcomes, the authors concluded that the risk of bias also differed.

In most systematic reviews, however, authors typically report a figure or table with risk of bias assessments for each study without distinguishing between outcomes. Although one might infer that in such instances authors considered the extent of risk of bias to be similar for each outcome, they may have failed to consider the possibility of different risk of bias across outcomes.Thus, an explicit statement from Core GRADE users that they did consider the issue and either concluded risk similar for all outcomes, or report separately for different outcomes, would be ideal.

3.4 Deciding Whether Individual RCTs or NRSI are at Low or High Risk of Bias

The extent of risk of bias in an individual study represents a continuum from minimal to extremely serious risk of bias. For simplicity, however, Core GRADE users can assess the overall risk of bias in individual studies as low or high. This judgment requires a threshold differentiating the two categories and the acknowledgment of close call situations (Figure 1). The arrows in Figure 1 remind us that risk of bias may be close to our threshold and that close call situations may bear on subsequent decisions.

Figure 1. Judging an individual study as overall high or low risk of bias



For example, consider the outcome of all-cause mortality in an RCT not using blinding in which randomization is concealed, follow-up is complete and there are no other concerns regarding risk of bias. The only important source of bias, co-interventions, arises from the lack of blinding of healthcare providers. Core GRADE users must then consider the likelihood of important co-intervention that may be highly impactful in one context (e.g. a heart failure trial with many potent treatments that may be differentially administered to intervention and control groups) versus low in another (e.g. multiple sclerosis, where few potent co-interventions exist and none with demonstrated impact on mortality). In the first context for the mortality outcome Core GRADE users would be likely to rate down for risk of bias due to lack of blinding, and in the second, users would be unlikely to do so. One might consider these and other similar situations as close call decisions regarding rating down the RCT for overall risk of bias.

Moreover, there is no definitive way to establish what the threshold should be regarding the number of high risk of bias items that merit rating a study as overall high risk of bias. One might do so for only one high risk category or item or require two or even more high risk categories or items to classify a study as high risk of bias. Thus, review teams may – and indeed do - use different thresholds.

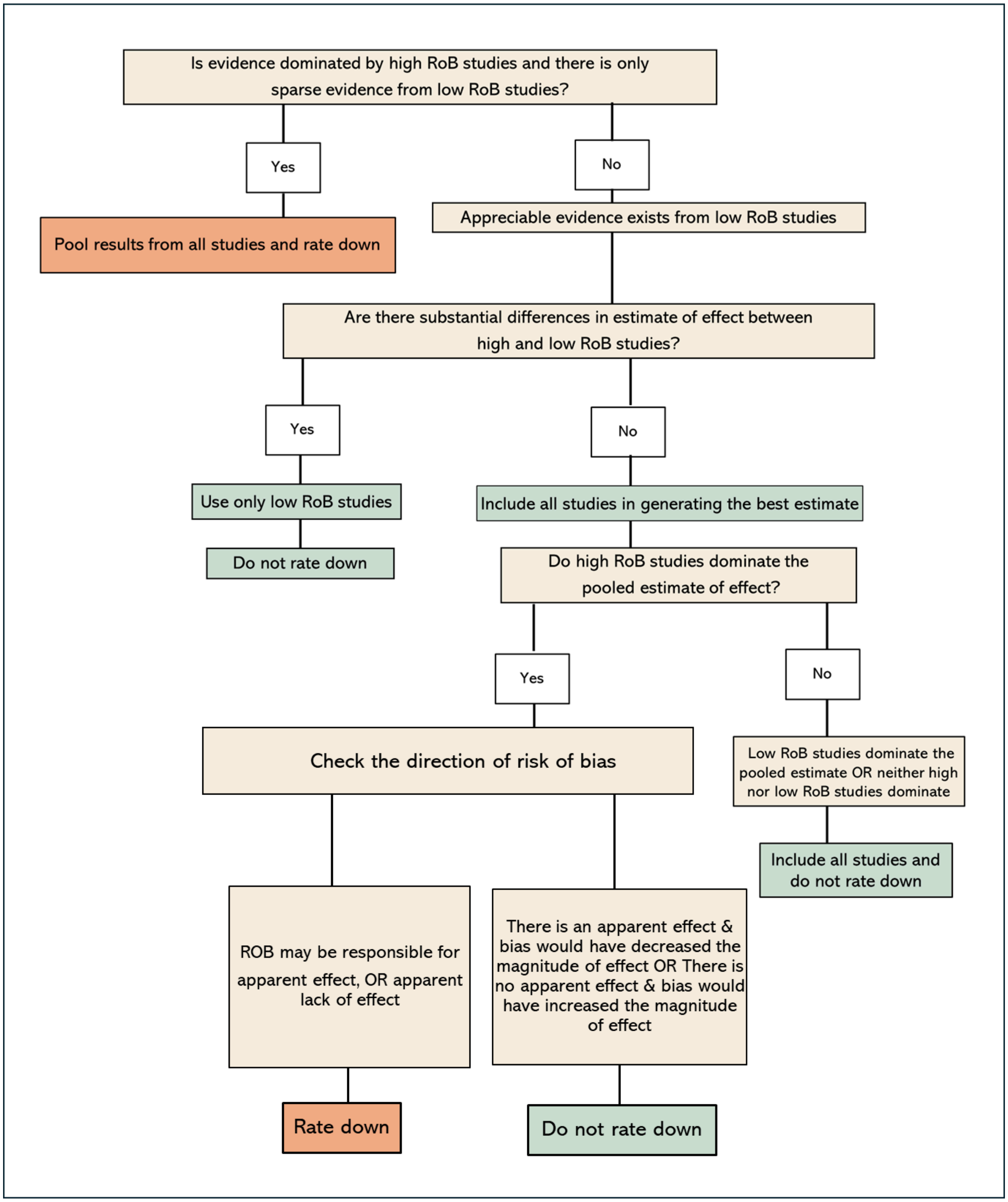
For example, in a systematic review of RCTs addressing the effect of gastrointestinal bleeding prophylaxis with proton pump inhibitors among patients who are critically ill, the authors used ROBUST-RCT to assess risk of bias.[36] Regarding the threshold of overall risk of bias in individual trials, if at least one item was rated as high risk of bias, authors considered the trial as overall high risk of bias. In contrast, the systematic review of cohort studies examining the association between red and processed meat consumption and cardiometabolic outcomes[37] used CLARITY's modified instrument to rate risk of bias in the included cohort studies and required two or more of the seven items (authors omitted one irrelevant item) rated as high risk of bias to consider the overall risk of bias as high. Finally, in another systematic review evaluating the association between use of antipsychotic drug and fracture risk,[38] for the included cohort studies the authors used CLARITY's modified instrument to rate their risk of bias and considered a study at overall high risk of bias only if three or more of the eight items were assessed as high risk of bias.

The choice of threshold – high risk of bias in only one or more than one item or category – may be an issue that will be impossible to resolve: how often does risk of bias actually result in bias? We do not know the answer and are unlikely ever to know. Moreover, the answer is likely to be context specific. Rating down a study for a single item or for two items assessed as high risk of bias would be reasonable; any more would be questionable. Explicit statement of the threshold choice, as well as the rationale, would increase transparency.

4. Rating Risk of Bias Across Bodies of Evidence

The ultimate goal of assessing risk of bias in individual studies is to inform whether to rate down the certainty of evidence across the entire body of evidence. In addressing risk of bias across all studies, Core GRADE users should follow the steps summarized in Figure 2. The flow chart applies to both RCTs that will start as overall high certainty evidence and NRSI that will start as low certainty evidence. For NRSI the rating down for risk of bias refers to moving from low to very low certainty evidence.

Figure 2. A flow chart depicting the process of considering ratings of low or high risk of bias in individual studies to arrive at a decision of whether to rate down for risk of bias


Is evidence dominated by high RoB studies and there is only sparse evidence from low RoB studies?
Yes
No
Pool results from all studies and rate down
Appreciable evidence exists from low RoB studies
Are there substantial differences in estimate of effect between high and low RoB studies?
Yes
No
Use only low RoB studies
Do not rate down
Include all studies in generating the best estimate
Do high RoB studies dominate the pooled estimate of effect?
Yes
No
Check the direction of risk of bias
Low RoB studies dominate the pooled estimate OR neither high nor low RoB studies dominate
Include all studies and do not rate down
ROB may be responsible for apparent effect, OR apparent lack of effect
There is an apparent effect & bias would have decreased the magnitude of effect OR There is no apparent effect & bias would have increased the magnitude of effect
Rate down
Do not rate down

1) First, systematic reviewers need to determine if there is only sparse evidence from low risk of bias studies. If that is the case, Core GRADE users should pool results from all studies and rate down the certainty of evidence for risk of bias.

   For example, a systematic review compared the addition of azithromycin to scaling and root planing versus not adding azithromycin in patients with chronic periodontitis.[39] For bleeding on probing within 3 months, in the four relevant trials, the only low risk of bias trial enrolled only 28 patients and reported a mean difference of 5.43 with a very wide 95% confidence interval (CI) of -8.96 to 19.82. The authors therefore included both the single low and the three high risk of bias trials and calculated a pooled estimate of -6.65

(95% CI -10.41 to -2.89). Because of the sparse evidence from the single low risk of bias trials, the authors appropriately rated the certainty down for risk of bias.

2) When appreciable evidence from low risk of bias studies exists, Core GRADE users should consider whether low and high risk of bias studies suggest similar or substantially different magnitudes of effect for each outcome of interest. This inquiry should include formal tests of subgroup differences between trials, acknowledging that such tests are often underpowered when there are only few trials.

   i) If low and high risk of bias studies suggest substantially different intervention effects, Core GRADE users will base inferences on only the low risk of bias studies as their best estimate of effect and not rate the certainty down for risk of bias.

      For example, a systematic review investigating the effect of corticosteroid therapy for patients hospitalized with community-acquired pneumonia addressed the outcome of duration of stay in hospital.[40] The investigators judged three trials with 1,288 patients at low risk of bias and six trials with 359 patients at high risk of bias. The authors conducted subgroup analysis based on risk of bias and found very different estimates from low risk of bias trials (mean difference -1.00 day, 95% -1.79 to -0.21) and high risk of bias studies (mean difference -4.41 days, 95% CI -7.65 to -1.17) (interaction p-value 0.045). Thus, the authors based their inferences only on low risk of bias trials and did not rate down the certainty for risk of bias.

   ii) If results are not importantly different in low and high risk of bias studies, Core GRADE users will include all studies in generating their best estimate of intervention effects and consider two scenarios:

      a) If high risk of bias studies dominate the body of evidence – that is, they carry substantially more weight than low risk of bias studies - Core GRADE users will generally rate down the certainty of evidence for risk of bias.

         For example, a systematic review compared child feeding interventions to no intervention in children aged five years and under.[41] The meta-analysis of 15 trials with 1,976 participants suggested increased vegetable consumption in the child feeding intervention group (standardized mean difference 0.44, 95% CI 0.24 to 0.65). Subgroup analysis suggested similar results in low (standardized mean difference 0.65, 95% CI 0.07 to 1.23; five trials with 507 participants) and high risk of bias trials (standardized mean difference 0.38, 95% CI 0.16 to 0.59; ten trials with 1,469 participants). Thus, the authors used the pooled estimate to make inference and since high risk of bias trials dominated the pooled estimate – enrolling 74% of the participants and carrying 69% of the weight – the authors rated the certainty of evidence down for risk of bias.

      b) If, in contrast, low risk of bias studies dominate the pooled estimate of effect, Core GRADE users will use all eligible studies in their pooled estimates and will not rate down for risk of bias.

         For example, in a systematic review evaluating effect of proton pump inhibitors on preventing clinically important gastrointestinal bleeding in critically ill patients,[36] low risk of bias studies (relative risk 0.50, 95% CI 0.25 to 0.99; five studies with 8,482 patients) and high risk of bias studies (relative risk 0.66, 95% CI 0.27 to 1.63; four studies with 597 patients) suggested similar results. Thus, the reviewers included all studies in generating the best estimate of effect (relative risk 0.51, 95% CI 0.34 to 0.76). Since low risk of bias studies dominated the overall pooled effect estimate (the weight of low risk of bias studies was 90%), the reviewers did not rate down the certainty of evidence for risk of bias.

c) Core GRADE users will encounter situations in which neither low nor high risk of bias studies dominate the pooled estimate, that is, low and high risk of bias studies carry similar weights. In these circumstances, Core GRADE users needn't rate down for risk of bias. The reason: *risk* of bias is just that, a risk that in a particular instance may or may not actually create bias. Investigators may fail to conceal randomization, fail to blind, or lose large numbers of patients to follow-up and still generate minimally biased results. Thus, if low and high risk of bias studies suggest similar results, one can reasonably infer that the high risk of bias studies have provided minimally biased estimates and use results from all studies and not rate down the certainty of evidence for risk of bias.

   For instance, a systematic review of RCTs compared effect of human or bovine colostrum to placebo in preterm infants.[42] For the outcome time to reach full feed, four studies with 131 participants proved at low risk of bias and two studies with 154 participants proved at high risk due to lack of allocation concealment. Low and high risk of bias studies suggested similar results (low risk of bias studies: weighted mean difference -4.19 days, 95% CI -9.40 to 1.03; high risk of bias studies: weighted mean difference -3.47 days, 95% CI -9.06 to 2.13; interaction p-value = 0.85). Thus, reviewers used the results from all studies as the best effect estimate (weighted mean difference -3.55 days, 95% CI -6.77 to -0.33) and did not rate down certainty of evidence for risk of bias.

4.1 Considering Direction of Bias

Core GRADE users should also consider the expected direction of any bias influencing results. If, in studies showing no difference in effect of interventions, identified bias would have increased differences between groups, one would reasonably infer that actual differences of effects must be smaller than those observed. Consideration of risk of bias would thus, if anything, reinforce the conclusion of no difference between groups and there would be no reason to rate down for risk of bias. Similarly, if effects in studies show a difference, but bias would have decreased that difference, the inference would be that the true relative effect is if anything larger than that observed. There would therefore be no reason to rate down for risk of bias.

Consider for instance a meta-analysis of RCTs addressing prevention of *Plasmodium falciparum* malaria transmission that compared the addition of primaquine to a prior regimen versus not adding the drug.[43] Results provided evidence that primaquine reduced, rather than increased adverse events (odds ratio 0.79, 95% CI 0.63 to 1.42). Lack of blinding suggested classifying the studies at high risk of bias. However, if clinicians know that a patient is receiving an additional medication, they would be more inclined to attribute symptoms to side effects from the medication than in patients not receiving the medication. If they were blinded, there could be no such differential attribution. Thus, bias from failure to blind would have led to an overestimation of adverse events ofwith primaquine. Considering this, the direction of bias increases our strength of inference that primaquine does not importantly increase adverse effects. Thus, review authors appropriately decided against rating down for risk of bias.

5. Publication Bias

Publication bias refers to the bias in the pooled estimate of effect that results from failure to publish studies based on their results, typically failure to publish negative studies.[44] Analyses of trials registered with institutional review boards have demonstrated selective non-publication of studies with negative or statistically non-significant results.[45 46] The effect has proved greater in NRSI than randomized trials.[47 48]

There are at least three causes of selective non-publication of studies with negative results. First, authors may fail to submit studies for publication because of a perception that journals will consider negative results uninteresting. Second, journal editors and their peer reviewers may indeed find negative results uninteresting and reject manuscripts on that basis. Third, for commercially funded studies, it is in the interest of funders motivated to maximize use of their product to suppress negative results and thus create an impression of larger than actual beneficial effects.

5.1 Commercial Funding

In one example of selective publication by manufacturers, a systematic review examining effect of reboxetine on acute treatment of major depression retrieved both published trials from databases and unpublished data from the manufacturer of reboxetine.[49] Results showed that published data overestimated the benefit of reboxetine by up to 115% when compared to placebo and 23% when compared to selective serotonin reuptake inhibitors.

In another example, a study investigated 74 antidepressant trials registered in Food and Drug Administration (FDA) and found selective publication based on results.[50] The FDA deemed results from 38 studies as positive, of which 37 were published. While, among the 36 trials with results deemed as negative or questionable, 22 were not published and 11 were published as positive.

Another study analyzed 400 randomly selected trials in the ClinicalTrials.gov website regarding their public disclosure of results.[51] It found that 118 trials (29.5%) failed to make their results public within four years of completion. Commercially funded trials (adjusted hazard ratio 0.49, 95% CI 0.36-0.66) were less likely to be published or were published later.

Because of the concern regarding the impact of industry sponsorship on selective publication, Core GRADE users should consider rating down for publication bias when the available studies are all small and industry sponsors have conducted most or all studies. For instance, a systematic review of flavonoids in patients with hemorrhoids that demonstrated impressive relative risk reductions in bleeding and pain identified 11 studies ranging in size from 40 to 234 all of which were industry sponsored.[52]

5.2 Avoiding Publication Bias: Comprehensive Search

Consideration of publication bias creates a unique problem for Core GRADE users: one is guessing at the presence of something that one cannot document. Systematic reviews with a less-than-comprehensive search may not locate studies published in non-indexed or non-English journals, or studies in registries (e.g. clinicaltrials.gov) or regulatory databases (e.g. FDA and European Medicines Agency),[53] thus raising the possibility of conducting searches from these sources. For example, a systematic review of leukotriene receptor antagonists for chronic urticaria identified 24 out of 34 relevant RCTs in Chinese.[54]

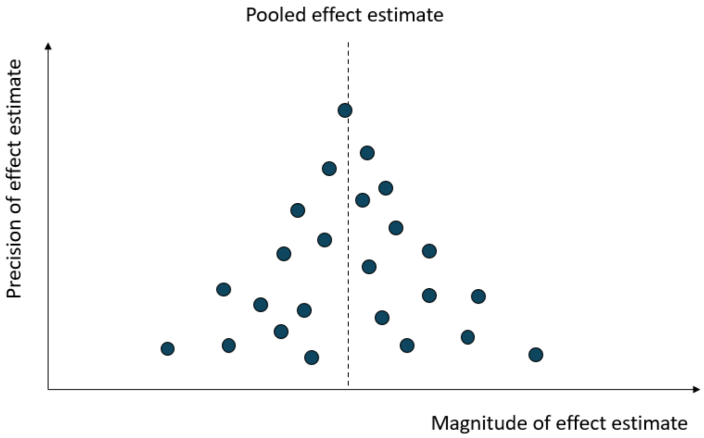
Nevertheless, the likely low yield precludes the necessity of such searches in all or even most cases. They may, however, be desirable in some instances such as searching for Chinese databases when conducting a systematic review of traditional Chinese medicine. Even a comprehensive search will not, however, detect studies with delay in publication, those that were never submitted, or do not appear in any study registries.[55]

5.3 Suspecting Publication Bias: Funnel Plots and Statistical Tests

Core GRADE users can assess risk of publication bias by visually inspecting the funnel plot, a scatter plot in which each dot represents a study included in the meta-analysis. The horizontal axis shows the magnitude of effect estimate of the individual studies (e.g., log odds ratio, mean difference), and the vertical axis shows precision of the estimate of effect (e.g., inverse of standard error, sample size).[56]

In a funnel plot, larger studies with more precise results are displayed at the apex and because they are more precise should be closer to the pooled estimate of effect. Smaller studies with lower precision scatter more widely at the bottom and should be symmetrically distributed around the pooled effect estimate. Thus, distribution of the dots should resemble a symmetrical inverted funnel (Figure 3a).

Figure 3a. Funnel plot that is not suggestive of publication bias



If the funnel plot is asymmetrical with a missing quadrant of small studies with negative results, publication bias represents a plausible explanation (Figure 3b). Other explanations include, however, small studies being biased in favor of the intervention, or small studies more faithfully following the intervention and thus achieving more favorable results. Given these alternative explanations we sometimes refer to such asymmetrical funnel plots as showing "small study effects".

Figure 3b. Funnel plot suggestive of publication bias

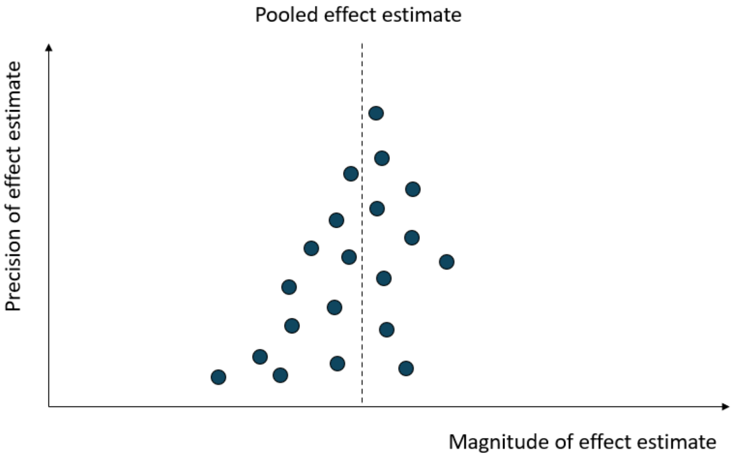
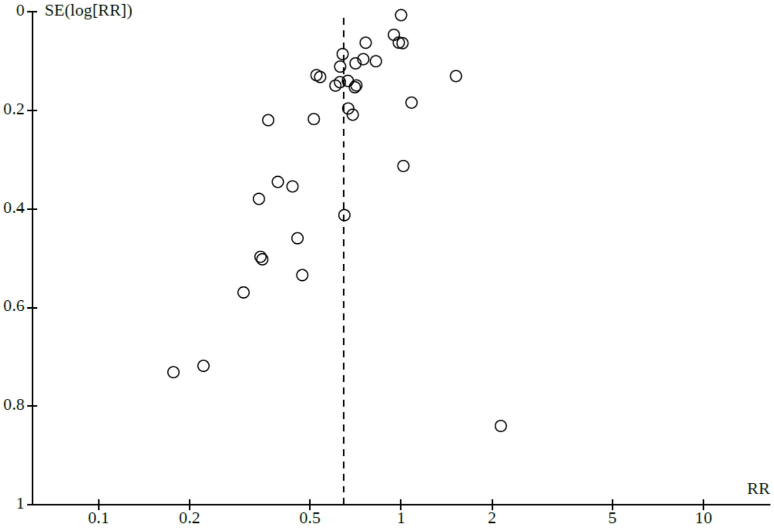


Figure 4 presents an example of funnel plot asymmetry from a systematic review investigating effects of probiotics on the risk of acute infectious diarrhoea lasting ≥ 48 hours,[57] in which there are a number of small studies favoring probiotics to a greater extent than the large studies but only one small study less favorable than the large studies. Such a result warrants serious consideration of rating down for publication bias.

Figure 4. Funnel plot in a systematic review investigating effects of probiotics on acute infectious diarrhoea suggested high risk of publication bias

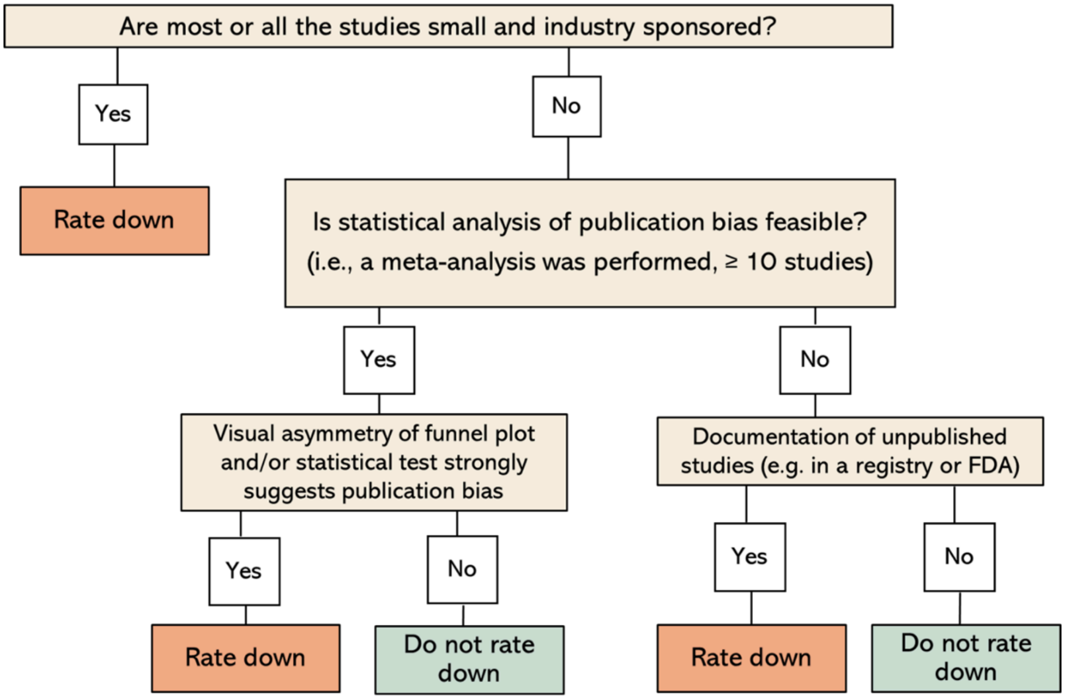


However, there are limitations of using funnel plots to test publication bias. Visual inspection of asymmetry involves subjectivity, which is prone to error.[58] Statistical approaches to test the asymmetry of funnel plot, including Egger's regression test,[59] Begg's rank test,[60] and a variety of other tests [61-64] are available but have been criticized for both false negative rates and false positive rates.[44 65] The use of statistical tests requires a meta-analysis including 10 or more studies, also preferable for making inferences regarding funnel plot asymmetry.[66]

Because of the limitations of the approaches for assessing publication bias, Core GRADE users will often be left with uncertainty. Core GRADE therefore suggests using the terms "undetected" (when no evidence suggests publication bias and they thus do not rate down certainty, the usual situation) and "strongly suspected" (when evidence suggesting publication bias exists and they do rate down certainty) to describe the publication bias domain.[55]

Figure 5 shows the steps Core GRADE users can follow to decide whether to rate down certainty of evidence for publication bias.

Figure 5. A flow chart depicting the process of deciding whether to rate down certainty of evidence for publication bias


Are most or all the studies small and industry sponsored?
Yes
No
Rate down
Is statistical analysis of publication bias feasible?
(i.e., a meta-analysis was performed, ≥ 10 studies)
Yes
No
Visual asymmetry of funnel plot and/or statistical test strongly suggests publication bias
Documentation of unpublished studies (e.g. in a registry or FDA)
Yes
No
Yes
No
Rate down
Do not rate down
Rate down
Do not rate down

6. Selective Outcome Reporting

One type of selective outcome reporting occurs when the results for an outcome of interest in some studies are unfavorable thus the investigators do not report the results. In this case, these studies do not contribute to the meta-analysis for that outcome. Since the funnel plot and test for funnel plot asymmetry can detect this problem, it is addressed in the publication bias domain.

There is another type of selective outcome reporting. The studies have reported the results for outcome of interest; thus, they have been included in meta-analysis. However, the reported result is selected from multiple available effect estimates (e.g., multiple time points, multiple outcome measurement methods, or multiple analytic approaches) or there is inconsistency between the protocol and the publication report in the outcome measurement. NRSI, which are typically far less rigorously prespecified with pre-registered protocols, require special attention since reporting bias may be enormous.[67 68] This type of selective outcome reporting should be addressed as risk of bias in individual studies rather than publication bias.

7. Rating Up Certainty of Evidence

While NRSI start out as low certainty evidence, it is possible to rate up certainty derived from NRSI to moderate or even high certainty. We will now review the two situations when Core GRADE users might rate up certainty: large magnitude of effect and dose-response gradient.

7.1 Large Magnitude of Effect

As we have described, in the Core GRADE approach NRSI start out at low certainty and may then be rated down for risk of bias issues particular to NRSI designs. When NRSIs were not rated down from low to very low (i.e. no risk of bias limitations particular to NRSI designs, and sufficiently precise to exclude values less extreme than our suggested thresholds), Core GRADE users will consider whether they show large effects. If they do observe large effects, Core GRADE users will consider rating up the certainty of evidence using the following thresholds: relative risk greater than 2.0 or less than 0.5 (similar thresholds for odds and hazard ratio), consider rating certainty up one level; when relative risk is greater than 5.0 or less than 0.2, consider rating up two levels.[69] The rationale for this guidance is that modeling studies have demonstrated that the likelihood of confounders that could explain a relative risk greater than 2.0 or less than 0.5 is low and the likelihood of confounders that could explain a relative risk greater than 5.0 or less than 0.2 is very low.[70]

For example, a systematic review of observational studies examining the relationship between infant sleeping position and sudden infant death syndrome (SIDS) found an odds ratio of 4.46 (95% CI 2.98 to 6.68) of SIDS occurring with front vs. back sleeping positions.[71] Such an association would warrant rating up certainty by one level.

Other factors may strengthen the case for rating up. These include rapidity of onset (e.g. insulin for diabetic ketoacidosis and epinephrine to treat anaphylaxis) and a relentless downhill trajectory without intervention (e.g. hip replacement for severe hip osteoarthritis).[69 72]

7.2 Dose-Response Gradient

The term dose-response gradient describes an observation that incremental increases (or decreases) of the exposure or intervention produce incremental increases (or decreases) in the effect. For example, a meta-analysis of salvage radiotherapy following radical prostatectomy demonstrated that each 1 Gy increase in the dose of radiotherapy is associated with a 2% increase in relapse-free survival.[73] This dose-response gradient increases our certainty that a causal connection between the intervention and the outcome exists.

There are, however, dangers in rating up for a dose-response gradient if the putative causal exposure is not actually causal, but linked to another exposure that is causal.[74] For example, several case-control studies showed a dose-response gradient between coffee consumption and pancreatic cancer.[75 76] As it turns out, the real culprit is not coffee but smoking, which does cause pancreatic cancer and for which a true dose-response gradient exists (the more you smoke, the higher the likelihood of pancreatic cancer).[77 78] The apparent dose-response gradient for coffee was a result of an association between smoking and coffee consumption: smokers drank more coffee, and the more they smoked, the more coffee they drank.[79 80] If Core GRADE users suspect such confounding between causal and non-causal associations, they will not rate up certainty for dose-response.[74]

Appendix 2 presents an alternative conceptualization of the rating up process in which studies with large or very large effects and/or a credible dose-response gradient begin at moderate or high certainty evidence.

8. Conclusion

Core GRADE users will start by assessing individual studies as being at low or high risk of bias. When considering all studies together to decide whether to rate down certainty of evidence for risk of bias, Core GRADE users will consider whether results differ importantly in studies at low and high risk of bias. If they do, Core GRADE users will use only the results from low risk of bias studies as best estimates of effect. If results of low and high risk of bias studies are similar, Core GRADE users will use the pooled results from all studies: if high risk of bias studies dominate, they will rate down for risk of bias; otherwise, they needn't rate down.

While all approaches to addressing publication bias have limitations, considering funnel plots or statistical approach for testing funnel plot asymmetry may be of use. A body of evidence consisting of relatively small studies funded by industry providers should raise suspicion. When magnitude of effect is large or very large, or a credible dose-response gradient exists, and the risk of bias is overall deemed low, one can consider rating up certainty of evidence from NRSI.

References

1. Wang Y, Ghadimi M, Wang Q, et al. Instruments assessing risk of bias of randomized trials frequently included items that are not addressing risk of bias issues. *J Clin Epidemiol* 2022;152:218-25. doi: 10.1016/j.jclinepi.2022.10.018 [published Online First: 20221028]
2. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928. doi: 10.1136/bmj.d5928 [published Online First: 20111018]
3. Sterne JAC, Savovic J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:l4898. doi: 10.1136/bmj.l4898 [published Online First: 20190828]
4. CLARITY Group. Tool to Assess Risk of Bias in Randomized Controlled Trials [Available from: https://www.distillersr.com/resources/methodological-resources/tool-to-assess-risk-of-bias-in-randomized-controlled-trials-distillersr.
5. Critical Appraisal Skills Programme (CASP). CASP randomised controlled trial standard checklist [Available from: https://casp-uk.net/casp-tools-checklists/randomised-controlled-trial-rct-checklist/.
6. Joanna Briggs Institute (JBI). Checklist for Randomized Controlled Trials [Available from: https://jbi.global/sites/default/files/2020-08/Checklist_for_RCTs.pdf.
7. National Institute for Health and Care Excellence (NICE). Methodology checklist: randomised controlled trials [Available from: https://www.nice.org.uk/process/pmg6/resources/the-guidelines-manual-appendices-bi-2549703709/chapter/appendix-c-methodology-checklist-randomised-controlled-trials.
8. Scottish Intercollegiate Guidelines Network (SIGN). Methodology checklist 2: randomized controlled trials [Available from: https://www.sign.ac.uk/using-our-guidelines/methodology/checklists/.
9. Wang Y, Keitz SA, Briel M, et al. Development of the Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT). Under review by BMJ.
10. Kuehn R, Wang Y, Guyatt G. Overly complex methods may impair pragmatic use of core evidence-based medicine principles. *BMJ Evid Based Med* 2024 doi: 10.1136/bmjebm-2024-112868 [published Online First: 20240307]
11. Moore THM, Higgins JPT, Dwan K. Ten tips for successful assessment of risk of bias in randomized trials using the RoB 2 tool: Early lessons from Cochrane. *Cochrane Evidence Synthesis and Methods* 2023;1(10):e12031. doi: https://doi.org/10.1002/cesm.12031
12. Crocker TF, Lam N, Jordao M, et al. Risk-of-bias assessment using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-intensive: observations from a systematic review. *J Clin Epidemiol* 2023;161:39-45. doi: 10.1016/j.jclinepi.2023.06.015 [published Online First: 20230624]
13. Minozzi S, Cinquini M, Gianola S, et al. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *J Clin Epidemiol* 2020;126:37-44. doi: 10.1016/j.jclinepi.2020.06.015 [published Online First: 20200618]
14. Wang Y, Parpia S, Couban R, et al. Compelling evidence from meta-epidemiological studies demonstrates overestimation of effects in randomized trials that fail to optimize randomization and blind patients and outcome assessors. *J Clin Epidemiol* 2024;165:111211. doi: 10.1016/j.jclinepi.2023.11.001 [published Online First: 20231107]
15. Miller AB, Goff DC, Bammann K, et al. Cohort Studies. In: Ahrens W, Pigeot I, eds. Handbook of Epidemiology: Springer New York, NY 2014:259-91.
16. Levine M, Walter S, Lee H, et al. Users' guides to the medical literature. IV. How to use an article about harm. Evidence-Based Medicine Working Group. *JAMA* 1994;271(20):1615-9. doi: 10.1001/jama.271.20.1615
17. Breslow NE. Case-control studies. In: Ahrens W, Pigeot I, eds. Handbook of Epidemiology: Springer New York, NY 2014:293-323.
18. D'Andrea E, Vinals L, Patorno E, et al. How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools. *BMJ Open* 2021;11(3):e043961. doi: 10.1136/bmjopen-2020-043961 [published Online First: 20210324]

19. Jiu L, Hartog M, Wang J, et al. Tools for assessing quality of studies investigating health interventions using real-world data: a literature review and content analysis. *BMJ Open* 2024;14(2):e075173. doi: 10.1136/bmjopen-2023-075173 [published Online First: 20240213]
20. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int J Epidemiol* 2007;36(3):666-76. doi: 10.1093/ije/dym018 [published Online First: 20070430]
21. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of non randomized studies in meta-analyses [Available from: https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.
22. CLARITY Group. Tool to Assess Risk of Bias in Cohort Studies [Available from: https://www.distillersr.com/resources/methodological-resources/tool-to-assess-risk-of-bias-in-cohort-studies-distillersr.
23. CLARITY Group. Tool to Assess Risk of Bias in Case Control Studies [Available from: https://www.distillersr.com/resources/methodological-resources/tool-to-assess-risk-of-bias-in-case-control-studies-distillersr.
24. Sterne JA, Hernan MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919. doi: 10.1136/bmj.i4919 [published Online First: 20161012]
25. JA S, J H. The Risk Of Bias In Non-randomized Studies – of Interventions, Version 2 (ROBINS-I V2). 2024 [Available from: https://sites.google.com/site/riskofbiastool/welcome/robins-i-v2 accessed Jan 1 2025.
26. Schunemann HJ, Cuello C, Akl EA, et al. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. *J Clin Epidemiol* 2019;111:105-14. doi: 10.1016/j.jclinepi.2018.01.012 [published Online First: 20180209]
27. Igelstrom E, Campbell M, Craig P, et al. Cochrane's risk of bias tool for non-randomized studies (ROBINS-I) is frequently misapplied: A methodological systematic review. *J Clin Epidemiol* 2021;140:22-32. doi: 10.1016/j.jclinepi.2021.08.022 [published Online First: 20210823]
28. Jeyaraman MM, Rabbani R, Copstein L, et al. Methodologically rigorous risk of bias tools for nonrandomized studies had low reliability and high evaluator burden. *J Clin Epidemiol* 2020;128:140-47. doi: 10.1016/j.jclinepi.2020.09.033 [published Online First: 20200925]
29. Minozzi S, Cinquini M, Gianola S, et al. Risk of bias in nonrandomized studies of interventions showed low inter-rater reliability and challenges in its application. *J Clin Epidemiol* 2019;112:28-35. doi: 10.1016/j.jclinepi.2019.04.001 [published Online First: 20190411]
30. Losilla JM, Oliveras I, Marin-Garcia JA, et al. Three risk of bias tools lead to opposite conclusions in observational research synthesis. *J Clin Epidemiol* 2018;101:61-72. doi: 10.1016/j.jclinepi.2018.05.021 [published Online First: 20180602]
31. Murad MH, Sultan S, Haffar S, et al. Methodological quality and synthesis of case series and case reports. *BMJ Evid Based Med* 2018;23(2):60-63. doi: 10.1136/bmjebm-2017-110853 [published Online First: 20180202]
32. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *J Clin Epidemiol* 2011;64(4):407-15. doi: 10.1016/j.jclinepi.2010.07.017 [published Online First: 20110119]
33. Rabeneck L, Paszat LF, Hilsden RJ, et al. Bleeding and perforation after outpatient colonoscopy and their risk factors in usual clinical practice. *Gastroenterology* 2008;135(6):1899-906, 906 e1. doi: 10.1053/j.gastro.2008.08.058 [published Online First: 20080913]
34. Chai-Adisaksopha C, Alexander PE, Guyatt G, et al. Mortality outcomes in patients transfused with fresher versus older red blood cells: a meta-analysis. *Vox Sang* 2017;112(3):268-78. doi: 10.1111/vox.12495 [published Online First: 20170220]

35. Vernooij RWM, Zeraatkar D, Han MA, et al. Patterns of Red and Processed Meat Consumption and Risk for Cardiometabolic and Cancer Outcomes: A Systematic Review and Meta-analysis of Cohort Studies. *Ann Intern Med* 2019;171(10):732-41. doi: 10.7326/M19-1583 [published Online First: 20191001]
36. Wang Y, Parpia S, Ge L, et al. Proton-Pump Inhibitors to Prevent Gastrointestinal Bleeding - An Updated Meta-Analysis. *NEJM Evid* 2024;3(7):EVIDoa2400134. doi: 10.1056/EVIDoa2400134 [published Online First: 20240614]
37. Zeraatkar D, Han MA, Guyatt GH, et al. Red and Processed Meat Consumption and Risk for All-Cause Mortality and Cardiometabolic Outcomes: A Systematic Review and Meta-analysis of Cohort Studies. *Ann Intern Med* 2019;171(10):703-10. doi: 10.7326/M19-0655 [published Online First: 20191001]
38. Papola D, Ostuzzi G, Thabane L, et al. Antipsychotic drug exposure and risk of fracture: a systematic review and meta-analysis of observational studies. *Int Clin Psychopharmacol* 2018;33(4):181-96. doi: 10.1097/YIC.0000000000000221
39. Khattri S, Kumbargere Nagraj S, Arora A, et al. Adjunctive systemic antimicrobials for the non-surgical treatment of periodontitis. *Cochrane Database Syst Rev* 2020;11(11):CD012568. doi: 10.1002/14651858.CD012568.pub2 [published Online First: 20201116]
40. Siemieniuk RA, Meade MO, Alonso-Coello P, et al. Corticosteroid Therapy for Patients Hospitalized With Community-Acquired Pneumonia: A Systematic Review and Meta-analysis. *Ann Intern Med* 2015;163(7):519-28. doi: 10.7326/M15-0715
41. Hodder RK, O'Brien KM, Wyse RJ, et al. Interventions for increasing fruit and vegetable consumption in children aged five years and under. *Cochrane Database Syst Rev* 2024;9(9):CD008552. doi: 10.1002/14651858.CD008552.pub8 [published Online First: 20240923]
42. Sadeghirad B, Morgan RL, Zeraatkar D, et al. Human and Bovine Colostrum for Prevention of Necrotizing Enterocolitis: A Meta-analysis. *Pediatrics* 2018;142(2) doi: 10.1542/peds.2018-0767 [published Online First: 20180710]
43. Yilma D. Efficacy and safety of single-dose primaquine to interrupt Plasmodium falciparum malaria transmission in pediatric patients compared to adults: A WWARN systematic review and individual patient data meta-analysis. Submitted for publication.
44. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 2000;53(11):1119-29. doi: 10.1016/s0895-4356(00)00242-0
45. Dwan K, Gamble C, Williamson PR, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLoS One* 2013;8(7):e66844. doi: 10.1371/journal.pone.0066844 [published Online First: 20130705]
46. Hopewell S, Loudon K, Clarke MJ, et al. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev* 2009;2009(1):MR000006. doi: 10.1002/14651858.MR000006.pub3 [published Online First: 20090121]
47. Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *JAMA* 1992;267(3):374-8.
48. Easterbrook PJ, Berlin JA, Gopalan R, et al. Publication bias in clinical research. *Lancet* 1991;337(8746):867-72. doi: 10.1016/0140-6736(91)90201-y
49. Eyding D, Lelgemann M, Grouven U, et al. Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *BMJ* 2010;341:c4737. doi: 10.1136/bmj.c4737 [published Online First: 20101012]
50. Turner EH, Matthews AM, Linardatos E, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008;358(3):252-60. doi: 10.1056/NEJMsa065779
51. Saito H, Gill CJ. How frequently do the results from completed US clinical trials enter the public domain?--A statistical analysis of the ClinicalTrials.gov database. *PLoS One* 2014;9(7):e101826. doi: 10.1371/journal.pone.0101826 [published Online First: 20140715]

52. Alonso-Coello P, Zhou Q, Martinez-Zapata MJ, et al. Meta-analysis of flavonoids for the treatment of haemorrhoids. *Br J Surg* 2006;93(8):909-20. doi: 10.1002/bjs.5378
53. Ladanie A, Ewald H, Kasenda B, et al. How to use FDA drug approval documents for evidence syntheses. *BMJ* 2018;362:k2815. doi: 10.1136/bmj.k2815 [published Online First: 20180710]
54. Rayner DG, Liu M, Chu AWL, et al. Leukotriene receptor antagonists as add-on therapy to antihistamines for urticaria: Systematic review and meta-analysis of randomized clinical trials. *J Allergy Clin Immunol* 2024;154(4):996-1007. doi: 10.1016/j.jaci.2024.05.026 [published Online First: 20240607]
55. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol* 2011;64(12):1277-82. doi: 10.1016/j.jclinepi.2011.01.011 [published Online First: 20110730]
56. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001;54(10):1046-55. doi: 10.1016/s0895-4356(01)00377-8
57. Collinson S, Deans A, Padua-Zamora A, et al. Probiotics for treating acute infectious diarrhoea. *Cochrane Database Syst Rev* 2020;12(12):CD003048. doi: 10.1002/14651858.CD003048.pub4 [published Online First: 20201208]
58. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *J Clin Epidemiol* 2005;58(9):894-901. doi: 10.1016/j.jclinepi.2005.01.006
59. Egger M, Davey Smith G, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315(7109):629-34. doi: 10.1136/bmj.315.7109.629
60. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994;50(4):1088-101.
61. Peters JL, Sutton AJ, Jones DR, et al. Comparison of two methods to detect publication bias in meta-analysis. *JAMA* 2006;295(6):676-80. doi: 10.1001/jama.295.6.676
62. Duval S, Tweedie R. A Nonparametric “Trim and Fill” Method of Accounting for Publication Bias in Meta-Analysis. *Journal of the American Statistical Association* 2000;- 95(- 449):- 98.
63. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med* 2001;20(4):641-54. doi: 10.1002/sim.698
64. Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med* 2006;25(20):3443-57. doi: 10.1002/sim.2380
65. Lin L, Chu H. Quantifying publication bias in meta-analysis. *Biometrics* 2018;74(3):785-94. doi: 10.1111/biom.12817 [published Online First: 20171115]
66. Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;343:d4002. doi: 10.1136/bmj.d4002 [published Online First: 20110722]
67. Wang Y, Pitre T, Wallach JD, et al. Grilling the data: application of specification curve analysis to red meat and all-cause mortality. *J Clin Epidemiol* 2024;168:111278. doi: 10.1016/j.jclinepi.2024.111278 [published Online First: 20240212]
68. Zeraatkar D, Cheung K, Milio K, et al. Methods for the Selection of Covariates in Nutritional Epidemiology Studies: A Meta-Epidemiological Review. *Curr Dev Nutr* 2019;3(10):nzz104. doi: 10.1093/cdn/nzz104 [published Online First: 20190917]
69. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol* 2011;64(12):1311-6. doi: 10.1016/j.jclinepi.2011.06.004 [published Online First: 20110730]
70. Bross ID. Pertinency of an extraneous variable. *J Chronic Dis* 1967;20(7):487-95. doi: 10.1016/0021-9681(67)90080-x
71. Gilbert R, Salanti G, Harden M, et al. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. *Int J Epidemiol* 2005;34(4):874-87. doi: 10.1093/ije/dyi088 [published Online First: 20050420]

72. Glasziou P, Chalmers I, Rawlins M, et al. When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007;334(7589):349-51. doi: 10.1136/bmj.39070.527986.68
73. King CR. The dose-response of salvage radiotherapy following radical prostatectomy: A systematic review and meta-analysis. *Radiother Oncol* 2016;121(2):199-203. doi: 10.1016/j.radonc.2016.10.026 [published Online First: 20161115]
74. Murad MH, Verbeek J, Schwingshackl L, et al. GRADE GUIDANCE 38: Updated guidance for rating up certainty of evidence due to a dose-response gradient. *J Clin Epidemiol* 2023;164:45-53. doi: 10.1016/j.jclinepi.2023.09.011 [published Online First: 20230929]
75. MacMahon B, Yen S, Trichopoulos D, et al. Coffee and cancer of the pancreas. *N Engl J Med* 1981;304(11):630-3. doi: 10.1056/nejm198103123041102
76. Clavel F, Benhamou E, Auquier A, et al. Coffee, alcohol, smoking and cancer of the pancreas: a case-control study. *Int J Cancer* 1989;43(1):17-21. doi: 10.1002/ijc.2910430105
77. Zou L, Zhong R, Shen N, et al. Non-linear dose-response relationship between cigarette smoking and pancreatic cancer risk: evidence from a meta-analysis of 42 observational studies. *Eur J Cancer* 2014;50(1):193-203. doi: 10.1016/j.ejca.2013.08.014 [published Online First: 20130919]
78. Molina-Montes E, Van Hoogstraten L, Gomez-Rubio P, et al. Pancreatic Cancer Risk in Relation to Lifetime Smoking Patterns, Tobacco Type, and Dose-Response Relationships. *Cancer Epidemiol Biomarkers Prev* 2020;29(5):1009-18. doi: 10.1158/1055-9965.Epi-19-1027 [published Online First: 20200212]
79. Treur JL, Taylor AE, Ware JJ, et al. Associations between smoking and caffeine consumption in two European cohorts. *Addiction* 2016;111(6):1059-68. doi: 10.1111/add.13298 [published Online First: 20160327]
80. Bjørngaard JH, Nordestgaard AT, Taylor AE, et al. Heavier smoking increases coffee consumption: findings from a Mendelian randomization analysis. *Int J Epidemiol* 2017;46(6):1958-67. doi: 10.1093/ije/dyx147

Contributors: Gordon Guyatt, Victor Montori, Thomas Agoritsas, Monica Hultcrantz and Alfonso Iorio conceived and designed the Core GRADE series. Gordon Guyatt, Ying Wang, Prashanti Eachempati, Alfonso Iorio, and M Hassan Murad 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.