Core GRADE 5: rating certainty of evidence—assessing indirectness

Gordon Guyatt,[1,2,3] Alfonso Iorio,[1,2] Hans De Beer,[4] Andrew Owen,[5] Thomas Agoritsas,[1,3,6] M Hassan Murad,[7] Ganesan Karthikeyan,[8,9] Carlos Cuello,[1] Manya Prasad,[10] Kevin Kim,[1,11] Dalal S Ali,[12] Arnav Agarwal,[1,2,3] Lars G Hemkens,[13,14,15] Liang Yao,[16] Monica Hultcrantz,[17,18] Jamie Rylance,[19] Derek K Chu,[1,2] Per Olav Vandvik,[3,20] Benjamin Djulbegovic,[21] Reem A Mustafa,[1,22] Linan Zeng,[23,24,25] Prashanti Eachempati,[3,26,27] Bram Rochwerg,[1,2] Kameshwar Prasad,[28,29] Victor M Montori,[30,31] Romina Brignardello-Petersen[1]

For numbered affiliations see end of the article

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

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

Cite this as: *BMJ* 2025;389:e083865 http://dx.doi.org/10.1136/bmj-2024-083865

Accepted: 21 March 2025

This fifth article in a seven part series presents the Core GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach to systematic reviews, clinical practice guidelines, and health technology assessments and addresses issues of indirect evidence. Guideline developers and health technology assessment practitioners must carefully specify the population, intervention, comparison, and outcome (PICO)—their target PICO—and consider the extent to which the best available evidence matches their target. When target and study PICOs differ substantially, studies provide indirect evidence and Core GRADE users may rate down the certainty of evidence as a result of this indirectness. Whether examining studies from a search for direct evidence or a deliberate search for indirect evidence, for each substantial difference between target and study PICO, Core GRADE users must judge the likelihood that magnitude of effects will differ substantially. The greater the likelihood of substantial differences the more advisable rating down for indirectness.

SUMMARY POINTS

- GRADE (Grading of Recommendations Assessment, Development and Evaluation) distinguishes between two types of indirectness: indirect comparisons from network meta-analyses and indirectness related to PICO (population, intervention, comparison, outcome) elements; the latter is the focus of Core GRADE
- Indirectness arises when a mismatch occurs between PICO elements in the clinical question (the target PICO) and PICOs of the studies constituting the best available evidence
- When certainty of direct evidence is low, Core GRADE users should consider a formal search for indirect evidence
- Possible mismatches between the target population and the studies' populations include differences in age, changes in population over time, and differences in a condition or disease
- Possible mismatches between the target intervention and comparator and those of the studies' include the intensity or duration of the intervention and a comparator that represents suboptimal care
- Possible mismatches between target and studies' outcomes include duration of follow-up and, of most concern, use of a surrogate outcome

This is the fifth paper in a series describing Core GRADE (Grading of Recommendations Assessment, Development and Evaluation), the essentials of applying GRADE to the conduct and evaluation of systematic reviews, clinical practice guidelines, and health technology assessments addressing the effects of interventions. In the previous articles we presented an overview of the Core GRADE process[1] and discussed specific aspects of imprecision,[2] inconsistency,[3] and risk of bias.[4] This paper deals with the issue of what Core GRADE refers to as indirectness, and presents in sequence: the two types of indirectness—indirect comparisons and indirectness related to population, intervention, comparison, and outcome (PICO) issues; the different importance of indirectness in systematic reviews, guidelines, and health technology assessments; how to distinguish indirectness from inconsistency; indirectness that arises in searching for direct evidence versus relevant indirect evidence; due attention to indirect evidence; and examples of the various sources of indirectness.

The information in this article will enable Core GRADE users to understand two types of indirectness—indirect comparisons and indirectness related to PICO issues; distinguish indirectness encountered in two scenarios—in the search for direct evidence (which typically does not warrant rating down for indirectness) and in the deliberate search for indirect evidence (which typically warrants considering the issue of indirectness); and identify and evaluate the extent of indirectness related to concerns that arise for each PICO element.

Two types of indirectness

Indirect comparisons

Previous GRADE guidance has used the term indirectness in two ways (fig 1).[5] In one, which we

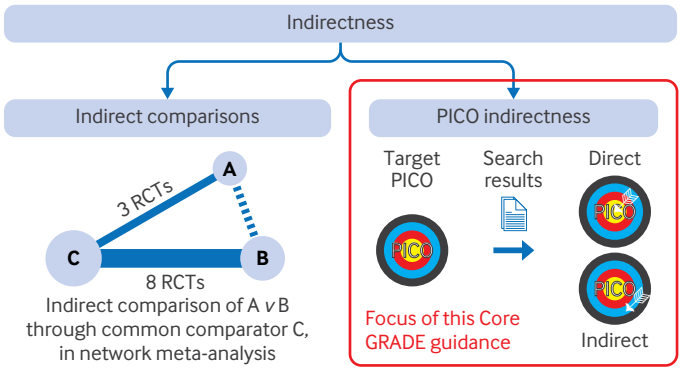

Fig 1 | Two types of GRADE indirectness. GRADE=Grading of Recommendations Assessment, Development and Evaluation; PICO=population, intervention, comparison, and outcome; RCT=randomised controlled trial

label indirect comparisons, the interest is in the relative merits of intervention A versus intervention B but evidence comes not from direct or head-to-head comparisons of A versus B but rather of A versus C and B versus C comparisons.[6] In this approach, for instance, if A does far better against C than does B, superiority of A over B is inferred.

Over the past 15 years, indirect comparisons have been almost entirely restricted to network meta-analyses that jointly consider multiple interventions and comparators. Core GRADE focuses on direct comparisons of a single intervention with a single comparator. This article will therefore not deal further with indirect comparisons but instead will focus on indirectness related to PICO issues.

Indirectness related to PICO issues

The Core GRADE approach begins with identifying a clinical question of interest and specifying the PICO. We refer to the clinical question of interest as the target PICO.

We define indirectness as a mismatch between the target PICO and the current best evidence. Research studies provide direct evidence for the population as enrolled, the intervention and comparison provided or used by the study participants, and outcomes as measured by investigators—the PICO elements in the study as carried out.

The study as carried out may not be the study as planned. Investigators may have sought a heterogenous population but enrolled only low risk patients, anticipated high adherence to the intervention and found only low adherence, anticipated one standard of care in the comparator but observed another, or planned a long follow-up but found that a funding shortfall necessitated a short follow-up. The mismatch between this direct evidence in the study as carried out and the target PICO can occur in any of the four elements of PICO. Henceforth, when we use the term indirectness, we will be referring to indirectness related to the target PICO rather than to indirect comparisons.

Indirectness concerns in guidelines and health technology assessments versus in systematic reviews

When researchers conduct systematic reviews independently from health technology assessments or guidelines, they can establish eligibility criteria that closely fit their target PICO and restrict their inclusion criteria accordingly. As a result, indirectness is not often a major concern in such reviews.

Health technology assessment practitioners and guideline developers must, in contrast, address questions of current interest to patients and clinicians. They choose, or are presented with, questions of sometimes urgent relevance to these target audiences. They must therefore identify and summarise the current best evidence to address those questions, even if that evidence represents a poor or limited match to their target PICO.

Indirectness versus inconsistency

Core GRADE users must attend to the possibility that intervention effects between their target PICO and available evidence will differ, requiring rating down for indirectness. On the one hand, if they have no reason to believe that relative effects differ between men and women, different drug doses, or outcomes measured over one versus three years, they will be unconcerned about applying results to women when most evidence comes from men, a higher dose when evidence comes from a lower dose, or three year outcomes when evidence comes from follow-up at one year. When, on the other hand, they believe relative effects are likely to differ, they will have concerns about indirectness.

We have dealt with this central issue—whether effects differ across subgroups of patients and interventions—in our overview of Core GRADE[1] and in our article addressing inconsistency.[3] In the latter, we noted that when systematic reviewers plan to use broad PICOs in their question definition (as should usually be the case), they must be prepared to face large differences in effects across studies. This preparation involves generating a priori hypotheses for possible explanations of inconsistency, and subsequently testing these hypotheses.

How do these issues of inconsistency and issues of indirectness differ? If we have evidence from both elderly people and younger people, low dose and high dose, or long follow-up and short follow-up, we can test whether effects differ across these variables. We label such situations as potential inconsistency and ultimately consider whether results suggest different effects between subgroups, and, if they do, evaluate the credibility of possible subgroup effects.[3 7]

However, if Core GRADE users are interested in effects in elderly people but all or almost all evidence comes from younger people, in low dose but all or almost all evidence comes from high dose, or in long follow-up but all or almost all evidence comes from short follow-up, they lack the data to test whether effects differ across these variables. Under these circumstances, they must use the indirect evidence from the younger people, the high dose, and the short follow-up to

make inferences about their target PICO. The extent to which relative effects will differ across such variables becomes a matter of mechanistic reasoning based on indirect evidence from basic research or other possibly analogous conditions, rather than on direct evidence from the patients and interventions under consideration. This method is thus less secure.

Indirectness in two scenarios

Encountered during search for direct evidence

A search for direct evidence sometimes yields evidence with some degree of indirectness involving one or more of the four PICO elements. Patients may be older or younger than the target population, have a different ethnic background, or have a different distribution of comorbidities.

Such differences typically do not warrant rating down for indirectness. The reason, as we have pointed out in previous papers in this series, is that true subgroup effects related to such characteristics are uncommon. Differences in baseline risk of adverse or desirable outcomes, including differences in comorbidity, seldom result in differences in relative effect.[8-12]

However, there are particular situations in which serious indirectness exists in studies that prove eligible in a search for direct evidence. Such situations include non-adherence to interventions, studies that focus on surrogate rather than on patient important outcomes, and problematic comparators.

When deliberately searching for indirect evidence

When direct evidence that matches their target PICO is unavailable or of very low or low certainty, Core GRADE users may fall back on evidence that substantially differs from their target PICO. When Core GRADE users deliberately search for indirect evidence, they will inevitably confront the possibility of rating down the certainty of evidence for indirectness.

Neglect of indirect evidence

Developers of clinical practice guidelines sometimes mistakenly conclude that no evidence exists for a PICO of interest. Very low quality evidence may, however, be available simply from clinical experience. Moreover, clinicians may often be considering an intervention because of evidence of its usefulness in related conditions—that is, indirect evidence. Consider, for instance, the repurposing of interventions at the onset of the covid-19 pandemic. The misguided enthusiasm for hydroxychloroquine[13] and ivermectin[14] highlights the limitations of such indirect evidence and thus the cautious inferences that it demands.

Indirect evidence may, even after rating down for indirectness has been considered, offer low or even moderate certainty. Even if such evidence offers only very low certainty, however, it remains preferable to making conclusions or decisions based on no evidence. Nevertheless, guideline developers sometimes neglect to consider indirect evidence. For example, a guideline panel may conclude there is no evidence for a potential intervention in children. They will often, however, be thinking exclusively of direct evidence. They may, as paediatricians would likely do in their clinical practice, be able to utilise indirect evidence from adults. In another example, early in the covid-19 pandemic, no direct evidence for several interventions existed, but indirect evidence from related conditions (eg, patients critically ill with acute respiratory distress syndrome but without covid-19) was available and provided support for guideline recommendations.[15]

Guideline developers who are not clear on the concept may use indirect evidence without explicitly labelling what they are doing. A study that evaluated guideline recommendations labelled as expert opinion found that most of these recommendations were in fact based on indirect evidence.[16]

Bearing in mind the possibility of indirect evidence, guideline developers and health technology assessment practitioners, when formulating search strategies for questions in which they anticipate sparse direct evidence, should seriously consider systematically searching for available indirect evidence that might inform their recommendations. Experts on the review team may be aware of the likelihood of finding relevant indirect evidence, and their advice may bear on the advisability of conducting the search.

Examples of indirectness: differences in population

Differences in age

Differences in age groups constitute a common indirectness issue in patients: elderly versus younger people, or children versus adults. For example, in a guideline that addressed the management of pancreatitis in children, authors found very limited evidence for antibiotic use in this age group. They therefore conducted a search for evidence from adults, ultimately using the indirect evidence as the basis for their recommendation. Although they did not conduct a formal certainty rating, authors described the evidence as limited, acknowledging decreased certainty associated with indirectness.[17]

Changes over time

Target patients may differ in many ways from patients enrolled in research studies. For example, the characteristics of the presenting patients may evolve over time, as occurred during the covid-19 pandemic. Casirivimab and imdevimab given in combination and sotrovimab given alone are monoclonal antibodies that bind to the SARS-CoV-2 spike protein, thus neutralising the virus. Randomised controlled trials conducted in 2020 and 2021 showed that both casirivimab and imdevimab combined and sotrovimab alone reduced mortality in patients infected with the circulating virus at that time, motivating World Health Organization recommendations for use of these agents.

However, changes in the sequence of the virus spike protein that occurred when omicron or its subsequent sublineages became the dominant variants resulted in substantially diminished neutralisation activity in vitro.[18] The population in the target PICO had

now changed from those infected with the viruses circulating earlier to those infected with the variants subsequently circulating. The panel had no direct evidence—that is, no studies of the antibodies in the era of the new virus variants were available. Nevertheless, the laboratory evidence of diminished antibody neutralisation led the panel to conclude that the original randomised controlled trials now provided only very indirect evidence for the key outcomes, the antibodies were very unlikely to be effective in the new target population, and strong recommendations against use of the antibodies were warranted.[19]

Similar challenges arise when, in searches for direct evidence, Core GRADE users must rely on results from older studies when diagnostic criteria and the availability of treatments differed. Relapsing and remitting multiple sclerosis provides an example of this phenomenon.[20]

Differences in condition

On occasion, when direct evidence is unavailable or of low or very low certainty, systematic review authors can look to populations with some similarity but nevertheless considerable differences from the target population. For instance, a review team addressed the choice of mechanical or bioprosthetic valves in patients with dialysis dependent end stage kidney disease who required surgery for valvular heart disease.[21] Patients receiving mechanical valves require long term anticoagulation whereas those receiving bioprosthetic valves do not. Observational studies comparing the two valve types provided only very low certainty evidence for one of the authors' key outcomes—postoperative and non-gastrointestinal bleeding at latest follow-up.

Given the very low certainty evidence, the authors sought indirect evidence and conducted a systematic review and meta-analysis of five randomised controlled trials of warfarin versus placebo in other populations. They found an incidence rate ratio for bleeding of 2.99 (95% confidence interval (CI) 1.46 to 6.13) which, after rating down for indirectness of the population, they considered moderate certainty evidence of increased bleeding with the mechanical heart valves.

Indirect evidence for harms

In rare conditions, randomised controlled trials are typically small or very small. Estimates of intervention harms may therefore yield very wide CIs warranting rating down twice for imprecision.

The interventions in such situations may have been repurposed after use in much larger populations with other conditions. Although it would be unwise to assume similar benefits across these conditions and the new indication, one might expect the adverse effects associated with a drug to be similar irrespective of the illness for which it is administered.

One might therefore rate down for indirectness only once—or not at all—for harms. Accordingly, if one had high certainty evidence for harms in other conditions, one would have moderate or high certainty for the population of immediate interest. Core GRADE users have applied these principles. Examples include the use of steroids in other inflammatory conditions to its use in thrombotic thrombocytopenic purpura[22] and chronic urticaria,[23] and allergen immunotherapy in asthma and allergic rhinitis to its use in atopic dermatitis.[24]

Systematic review authors have also applied the same principle to related conditions to improve the precision (ie, narrow CIs) of the estimates of harms across each of these conditions. For instance, a systematic review team pooled data from trials of corticosteroid use in sepsis, acute respiratory distress syndrome, and community acquired pneumonia to generate precise estimates of adverse effects.[25]

Examples of indirectness: differences in interventions

Interventions studied may differ from the target PICO in several ways, including dose of a drug (higher or lower than the target intervention), duration of administration (shorter or longer), route of administration (parenteral versus oral), or the skill level of providers of interventions such as in educational, surgical, physiotherapy, and psychosocial interventions. Another concerning common source of indirectness for such interventions is that authors may not sufficiently describe the components of the interventions and this failure can preclude their replication. For instance, the details for cardiac rehabilitation were so poorly reported in the literature that surveys of rehabilitation programmes showed that what they implemented in practice differed substantially from what randomised controlled trials had shown to be effective.[26] Inadequate description of the intervention constitutes a reason for rating down for indirectness.

Non-adherence

Another common way that trials of interventions differ from the target interventions is non-adherence of patients. Generally, patients and their healthcare providers are interested in the impact of an intervention when used as intended. High levels of non-adherence introduce problematic indirectness and thus compromise the certainty of the evidence.

For example, a randomised controlled trial of nortriptyline as an adjunct to nicotine replacement for smoking cessation randomised 901 adults attending a smoking cessation service to nortriptyline or placebo.[27] They found that one year after quit day, 11% in the nortriptyline group versus 9% in the control group (relative risk 1.26, 95% CI 0.84 to 1.87) had stopped smoking. However, much earlier, four weeks after quit day, only 59% of patients in the treatment group and 56% of patients in the control group were taking the drugs. Had adherence been close to 100%, the impact of the intervention may have been greater, the estimate more precise, and the evidence would warrant higher certainty. The trial thus provides only indirect evidence of the effect of nortriptyline on

smoking cessation in those who use the intervention.[28] In a systematic review that included additional trials that also had concerns about adherence, the CI was narrower (relative risk 1.29, 95% CI 0.97 to 1.72) suggesting that nortriptyline may increase smoking cessation (low certainty evidence due to indirectness and imprecision).[29]

Indeed, the indirectness here is serious enough that, if the target PICO specified the effects of the intervention when people use it, the extent of non-adherence would surely warrant rating down for indirectness. Even though adherence was very limited, results suggested a possible signal in favour of nortriptyline. It is entirely plausible that had adherence been very high the results would have shown a benefit of nortriptyline in improving quit rates in smokers.

Note that if the intervention of the target PICO included how nortriptyline was actually used in the community, one might conclude that the low adherence study provided direct evidence. Such targets that include considerations at the population or public health level occur particularly often in health technology assessments.

In trials of screening interventions, when the target PICO may well focus on those who are adherent, indirectness due to those who are not adherent is often a major problem. Consider a randomised trial of colonoscopy screening for colorectal cancer versus no such screening, and a PICO of interest that specifies that patients all undergo the screening intervention. The question of interest to patients would be: "What will be the impact if I undergo screening?"—and contrasts with the question of interest to the policy maker: "What will be the effect of instituting a programme in which only some of the eligible population will be interested?"

A randomised controlled trial of colonoscopy allocated more than 84 000 participants in Norway (highest participation >60%), Poland (lowest participation 33%), and Sweden to receive or not receive an invitation for colorectal screening. In the intention-to-screen analysis, the intervention reduced the relative risk of developing colorectal cancer by 18% (95% CI 7% to 30%). The absolute risk reduction was about 2 in 1000 population over 10 years—a magnitude of effect some might consider not worth the burden of screening. The evidence is, however, indirect: the investigators would presumably have seen a larger effect if all those invited had participated.

Indeed, a per protocol analysis focusing on the Norwegian population estimated that, if adherent, patients would experience a 45% relative risk reduction. With this estimate, the effect is still small but appreciably greater, about 6 per 1000 population. The per protocol analysis provides a more direct estimate but with increased risk of bias. In general, the greater the differences between the anticipated effect in a fully adherent population compared with the effect observed in the partially adherent population studied, the more likely Core GRADE users will rate down for indirectness.

Finally, it is possible that randomised trials in which patients achieved high adherence may provide indirect evidence from a public health or funder's point of view. Studies of behavioural interventions that most patients find extremely challenging to follow may enroll particularly committed patients and implement adherence enhancing strategies that are unfeasible or not widely applicable. They may thus achieve adherence that is unrealistically high for clinical practice.[30 31] From a public health point of view, putting resources into such interventions for typical patients who cannot achieve high levels of adherence may be a poor decision. The high adherence situation thus represents, from the policy makers' perspective, problematic indirect evidence.

Trials that allow switching treatments

Oncology trials may have protocols that allow switching treatments when a patient does not respond to the original intervention. For instance, consider the relative effects of two anticancer drugs, interferon-alfa and sunitinib, in adults with renal cancer. Systematic review authors encountered trials in which participants who experienced disease progression after treatment with interferon-alfa received sunitinib and other related treatments.[32] How might this design bear on issues of indirectness?

The answer lies in considering the target PICO.[33] Core GRADE users whose target PICO designates the comparison of sunitinib alone to interferon-alfa including the proviso that patients who do not respond to the drug are offered sunitinib will find results directly applicable. Systematic review authors interested in the impact of the two drugs without such switching will, in contrast, face limitations in the directness of the results.

In the latter case in which users are interested in the independent effect of the drugs, the extent of indirectness will depend on the proportion of participants in the intervention arm who switched to the alternative intervention. If the proportion of patients who switched is large, the indirectness may be considerable and warrant rating down. If few patients switched, indirectness may be minimal and not warrant rating down.

A second determinant of the necessity to rate down would be the apparent effect of the interventions. Considering the example comparison, if substantial switching to sunitinib occurs and the result for the arm that began with interferon-alfa proves similar to that of the sunitinib arm, the issue is in doubt: is the "rescue" sunitinib responsible for the similar results, or would the results have been achieved with interferon-alfa alone? On the other hand, if sunitinib proves superior, that superiority would only have been greater had no switching occurred. In the relevant systematic review, sunitinib and other related target treatments proved superior to interferon-alfa (relative survival 1.3, 95% CI 1.1 to 1.5). Thus, indirectness does not compromise the conclusion about sunitinib's superiority to interferon-alfa, and authors have no need to rate down for indirectness.

Change of intervention technology

When the intervention is a device or technology, its evolution over time can result in important indirectness that lowers certainty. For example, devices that help people manage their diabetes are constantly changing. Continuous glucose monitoring systems were approved by the Food and Drug Administration in the late 1990s and have quickly evolved with new sensor technology such that wear time has lengthened from a few days to weeks and months. "Real-time" systems, systems managed with smart phones, and systems linked to insulin delivery pumps (closed loop systems) are now available. Guidelines on diabetes technology struggled with indirectness of older evidence and have continuously balanced two strategies: excluding studies of obsolete systems versus including studies of older systems and lowering certainty due to indirectness.[34-36]

Examples of indirectness: differences in comparators

Situations in which the comparator differs from that in the target PICO include variations in standard care between jurisdictions, use of placebo when an active treatment is the clinically relevant active comparator, inferior older alternatives rather than current optimal alternatives, and differences in dose or route of administration.[37] These problems may arise in searches for direct evidence when systematic review authors do not explicitly identify their comparator.

The problematic use of placebos rather than active comparators is common,[38] particularly in drug development trials that have the ultimate goal of obtaining regulatory approval. For example, many randomised trials of disease modifying biologics for patients with rheumatoid arthritis did not use active comparators,[39] including trials enrolling patients with a high level of active disease, thus withholding potentially helpful treatments. While meeting regulatory requirements, such designs, by choosing suboptimal comparators, raise issues of indirectness.

The use of suboptimal comparators in industry sponsored trials is common, and include the following examples. Large industry sponsored trials evaluating newer antihypertensive drugs chose the beta blocker atenolol as the comparator, despite previous evidence showing inferiority of beta blockers to a low dose thiazide diuretic.[40] Manufacturers of newer antipsychotic agents overestimated the advantages of reduced toxicity of their drugs by comparing them to inappropriately large doses of older alternatives.[41] Eight such trials used fixed doses of haloperidol 20 mg/day, substantially above recommended doses.[42] Several studies used interferon beta-1a given intramuscularly as the comparator versus new drugs for multiple sclerosis after investigators had established the superiority of subcutaneous interferon alfa-2b.[43-46]

A more recent example comes from randomised controlled trials in patients with multiple myeloma conducted in the US in which enrolment occurred between 2010 and 2020.[47] The authors considered a control group regimen inferior if, before patient enrolment began, a previous randomised controlled trial had shown an improved progression-free survival versus the control group. Of 49 identified randomised controlled trials, seven (14%) began enrolling patients into inferior control groups after a study of an existing superior regimen had been published. The primary funding source in all seven was the pharmaceutical industry. These trials provide only indirect evidence for what might happen had trial investigators chosen the best available comparator. In 2000, a similar analysis of multiple myeloma trials illustrated the persistence of problems related to the selection of an inferior comparator.[48]

Chinese investigators studying randomised trials of anticancer drugs authorised by Chinese institutional review boards between 2016 and 2021 reported a similar problem. They found that 60 (13.2%) of 453 phase 2/3 and phase 3 randomised controlled trials included a suboptimal control arm.[49] In all these situations Core GRADE users would rate down the certainty of evidence for indirectness against the appropriate optimal comparator.

Investigators may sometimes have no choice but to use placebo comparisons to obtain indirect estimates of effects of alternative active agents. For instance, systematic review authors informing a clinical practice guideline were interested in interventions for the management of patients with X linked hypophosphataemia.[50] In particular, they wanted to evaluate the impact of burosumab on pain and function, both against no specific treatment and against conventional treatment of phosphate salts and active vitamin D. The authors identified a randomised controlled trial of burosumab versus placebo that provided moderate to high certainty evidence for some of the key outcomes, but no study comparing the drug with standard of care. They offered evidence from the trial against placebo as the best estimates representing the maximum differences against standard of care, rating down once for indirectness for each outcome. Although not a satisfactory situation, the authors approach is the best possible under the circumstances.

Differences in outcomes

The impact of intervention versus comparators on outcomes may differ as a result of how the outcomes are measured (eg, symptomatic versus asymptomatic vertebral fracture or symptomatic versus asymptomatic deep vein thrombosis), or the duration of follow-up (short term versus long term). Outcomes may also differ by how they are measured: directly (death rates) or indirectly through surrogate measures (reduction in viral load in HIV). Such issues will arise when Core GRADE users include studies that measure only surrogates and not patient important outcomes.

How should Core GRADE users handle the situation when the available outcome is a surrogate or substitute for what patients consider important? Core GRADE users will specify the patient important outcome for which the surrogate is substituting and consider the degree of indirectness, inferring the impact on the

Table 1 | Summary of findings table addressing long term heart failure symptoms in randomised controlled trials of percutaneous versus surgical mitral commissurotomy

Outcome	No of trials/No of patients	Result	Certainty of evidence
Heart failure symptoms as inferred from mitral valve area	Six randomised controlled trials, 458 patients	Little or no difference in symptoms of heart failure inferred from difference in mitral valve area of 0.13 cm^2 higher (95% CI 0.09 lower to 0.35 higher) in patients undergoing commissurotomy	Very low due to serious indirectness, serious imprecision, and serious inconsistency

patient important outcome from the surrogate and rating down certainty of evidence as appropriate. The following example provides an application of the approach.

Consider patients with mitral valve stenosis faced with the choice of percutaneous versus surgical mitral commissurotomy. A key outcome for such patients is progression of heart failure symptoms as a result of the procedure failing over the long term. Because patients with larger valve areas generally have fewer symptoms, studies comparing these procedures report the mitral valve area as a measure of success. A systematic review of randomised trials comparing the two procedures addressed their relative merits for minimising development or progression of heart failure symptoms.[51]

The review found that no eligible studies measured patient symptoms over the long term. What investigators conducting these studies did measure was a surrogate for symptoms—mitral valve area at 30 months by echocardiography or cardiac catheterisation. The systematic review authors specified their outcome of interest as patient symptoms over the long term as inferred from the surrogate. Ultimately, they rated down the certainty of evidence for imprecision and inconsistency as well as for indirectness of the outcome, resulting in very low certainty evidence as shown in an adaptation of their summary of finding table (table 1).

Although one might consider rating down more than one level for indirectness for any PICO element, this possibility is typically more salient for surrogate outcomes. For instance, in patients with end stage kidney disease, disturbances in calcium and phosphate metabolism may result in fragility fractures and myocardial infarction. Initial evidence of new therapeutic interventions focused on measures of calcium/phosphate metabolism, a very indirect measure of fractures and myocardial infarction, thus warranting rating down two levels for indirectness. Bone density for fractures and coronary calcification for myocardial infarction represent surrogates that may be better predictors of the impact of treatment on patient important adverse outcomes and thus may warrant rating down for indirectness by only one level.[51] Thus, the decision to rate down one or two levels depends on one's understanding of the likelihood that change in the patient important outcome will follow change in the surrogate.

GRADE users addressing the possible impact that treatment effects on surrogate outcomes might have on patient important outcomes might consider exercises in which they make specific quantitative assumptions and model likely results.[52] Because we see the likely gain in rigour and in soundness of conclusions as marginal, we do not see this as part of Core GRADE.

Table 2 summarises issues in rating down for indirectness, referring back to the examples we have used and presenting the likelihood of rating down.

Conclusion

Limitations in the extent to which the PICO in the available studies differs from the target PICO—in GRADE called indirectness—represent a common reason for rating down certainty of evidence in the development of guidelines and health technology assessments. When direct evidence is unavailable or of low or very low certainty, Core GRADE users should

Table 2 | Summary of indirectness issues

PICO element	Reason for rating down	Examples	Likelihood of rating down
Population	Population differences may interact with magnitude of effect	Adult versus paediatric Changes in virus antigens Comorbidity (diabetes, renal disease)	Low likelihood because relative effects are typically similar across populations
Intervention	Interventions often differ in dose, duration, or subclass	Drugs within a class Dose of drug Non-adherence Switching versus non-switching of treatments within a treatment strategy Advances in technology	Intermediate likelihood depending on underlying biology and on magnitude of issues such as non-adherence and frequency of switching
Comparison	Different comparators may have different effects on target outcomes	Use of placebos versus unblinded standard treatment or alternative intervention Inferior older alternatives in trials of new drug Suboptimal doses of comparators	Substantial likelihood in trials of new agents when an effective treatment already exists, particularly more than one effective treatment
Outcome	Impact on surrogates often fails to translate into improvement in patient important outcomes	Cardiac function versus mortality in heart failure Bone density versus fractures in osteoporosis Test performance versus function in dementia Blood glucose verus microvascular and macrovascular morbidity and mortality in diabetes	High likelihood because of frequent disappointing results in randomised controlled trials examining examining patient important outcomes

consider searching for indirect evidence that may result in higher certainty evidence. Whenever the PICO elements in the relevant studies do not completely correspond with Core GRADE users' target PICO, they must consider the likelihood that these differences will result in important variation in intervention effects, and if that is likely they should rate down by one level for indirectness or—particularly with surrogate outcomes—by two levels.

AUTHOR AFFILIATIONS

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

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

[3]MAGIC Evidence Ecosystem Foundation, Oslo, Norway

[4]Guide2Guidance, Lemelerberg 7, Utrecht, Netherlands

[5]Department of Pharmacology and Therapeutics, Centre of Excellence in Long-acting Therapeutics (CELT), University of Liverpool, Liverpool, UK

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

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

[8]Translational Health Science Technology Institute, Faridabad, India

[9]All India Institute of Medical Sciences, New Delhi, India

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

[11]Population Health Research Institute, Hamilton, ON, Canada

[12]Divisions of Endocrinology and Metabolism, McMaster University, Hamilton, ON, Canada

[13]Pragmatic Evidence Lab, Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland

[14]Department of Clinical Research, University Hospital Basel and University of Basel, Basel, Switzerland

[15]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA

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

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

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

[19]Liverpool School of Tropical Medicine, Liverpool, UK

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

[21]Division of Hematology/Oncology, Department of Medicine, Medical University of South Carolina, Charleston, SC, USA

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

[23]Pharmacy Department/Evidence-based Pharmacy Centre/ Children's Medicine Key Laboratory of Sichuan Province, West China Second University Hospital, Sichuan University, Chengdu, China

[24]Sichuan University and Key Laboratory of Birth Defects and Related Disease of Women and Children, Ministry of Education, Chengdu, China

[25]West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China

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

[27]Faculty of Dentistry, Manipal University College Malaysia, Malaysia

[28]Department of Neurology, All India Institute of Medical Sciences, New Delhi, India

[29]Fortis CSR Foundation, New Delhi, India

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

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

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

Funding: No external funding.

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

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

1 Gyuatt G, Agoritsas T, Brignardello-Petersen R, et al. Core GRADE 1: overview of the Core GRADE approach. *BMJ* 2025;389:e081903.
2 Guyatt G, Zeng L, Brignardello-Petersen R, et al. Core GRADE 2: choosing the target of certainty rating and assessing imprecision. *BMJ* 2025;389:e081904.
3 Guyatt G, Schandelmaier S, Brignardello-Petersen R, et al. Core GRADE 3: rating certainty of evidence—inconsistency. *BMJ* 2025;389:e081905.
4 Guyatt G, Wang Y, Eachempati P, et al. Core GRADE 4: rating certainty of evidence—risk of bias, publication bias, and reasons for rating up certainty. *BMJ* 2025;389:e083864.
5 Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;64:401-6. doi:10.1016/j.jclinepi.2010.07.015
6 Guyatt GH, Oxman AD, Kunz R, et al, GRADE Working Group. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol* 2011;64:1303-10. doi:10.1016/j.jclinepi.2011.04.014
7 Schandelmaier S, Briel M, Varadhan R, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. *CMAJ* 2020;192:E901-6. doi:10.1503/cmaj.200077
8 Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2002;21:1575-600. doi:10.1002/sim.1188
9 Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med* 1998;17:1923-42. doi:10.1002/(SICI)1097-0258(19980915)17:17<1923::AID-SIM874>3.0.CO;2-6
10 Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. *Int J Epidemiol* 2002;31:72-6. doi:10.1093/ije/31.1.72
11 Torres Roldan VD, Ponce OJ, Urtecho M, et al. Understanding treatment-subgroup effect in primary and secondary prevention of cardiovascular disease: An exploration using meta-analyses of individual patient data. *J Clin Epidemiol* 2021;139:160-6. doi:10.1016/j.jclinepi.2021.08.006
12 Hanlon P, Butterly EW, Shah AS, et al. Treatment effect modification due to comorbidity: Individual participant data meta-analyses of 120 randomised controlled trials. *PLoS Med* 2023;20:e1004176. doi:10.1371/journal.pmed.1004176
13 Martins-Filho PR, Ferreira LC, Heimfarth L, Araújo AAS, Quintans-Júnior LJ. Efficacy and safety of hydroxychloroquine as pre-and post-exposure prophylaxis and treatment of COVID-19: A systematic review and meta-analysis of blinded, placebo-controlled, randomized clinical trials. *Lancet Reg Health Am* 2021;2:100062. doi:10.1016/j.lana.2021.100062
14 Marcolino MS, Meira KC, Guimarães NS, et al. Systematic review and meta-analysis of ivermectin for treatment of COVID-19: evidence beyond the hype. *BMC Infect Dis* 2022;22:639. doi:10.1186/s12879-022-07589-8
15 Ye Z, Wang Y, Colunga-Lozano LE, et al. Efficacy and safety of corticosteroids in COVID-19 based on evidence for COVID-19, other coronavirus infections, influenza, community-acquired pneumonia and acute respiratory distress syndrome: a systematic review and meta-analysis. *CMAJ* 2020;192:E756-67. doi:10.1503/cmaj.200645
16 Ponce OJ, Alvarez-Villalobos N, Shah R, et al. What does expert opinion in guidelines mean? a meta-epidemiological study. *Evid Based Med* 2017;22:164-9. doi:10.1136/ebmed-2017-110798
17 Abu-El-Haija M, Kumar S, Quiros JA, et al. Management of Acute Pancreatitis in the Pediatric Population: A Clinical Report From the North American Society for Pediatric Gastroenterology, Hepatology and Nutrition Pancreas Committee. *J Pediatr Gastroenterol Nutr* 2018;66:159-76. doi:10.1097/MPG.0000000000001715

18 Arora P, Kempf A, Nehlmeier I, et al. Augmented neutralisation resistance of emerging omicron subvariants BA.2.12.1, BA.4, and BA.5. *Lancet Infect Dis* 2022;22:1117-8. doi:10.1016/S1473-3099(22)00422-4
19 Agarwal A, Hunt B, Stegemann M, et al. Therapeutics and COVID-19: living guideline 2023. https://app.magicapp.org/#/guideline/6989. [Accessed 20 Sept 2024.]
20 Zhang Y, Salter A, Wallström E, Cutter G, Stüve O. Evolution of clinical trials in multiple sclerosis. *Ther Adv Neurol Disord* 2019;12:1756286419826547. doi:10.1177/1756286419826547
21 Kim KS, Belley-Côté EP, Gupta S, et al. Mechanical versus bioprosthetic valves in chronic dialysis: a systematic review and meta-analysis. *Can J Surg* 2022;65:E450-9. doi:10.1503/cjs.001121
22 Zheng XL, Vesely SK, Cataland SR, et al. ISTH guidelines for the diagnosis of thrombotic thrombocytopenic purpura. *J Thromb Haemost* 2020;18:2486-95. doi:10.1111/jth.15006
23 Chu AWL, Rayner DG, Chu X, et al. Topical corticosteroids for hives and itch (urticaria): Systematic review and Bayesian meta-analysis of randomized trials. *Ann Allergy Asthma Immunol* 2024;133:437-444.e18. doi:10.1016/j.anai.2024.06.003
24 Yepes-Nuñez JJ, Guyatt GH, Gómez-Escobar LG, et al. Allergen immunotherapy for atopic dermatitis: Systematic review and meta-analysis of benefits and harms. *J Allergy Clin Immunol* 2023;151:147-58. doi:10.1016/j.jaci.2022.09.020
25 Chaudhuri D, Israelian L, Putowski Z, et al. Adverse Effects Related to Corticosteroid Use in Sepsis, Acute Respiratory Distress Syndrome, and Community-Acquired Pneumonia: A Systematic Review and Meta-Analysis. *Crit Care Explor* 2024;6:e1071. doi:10.1097/CCE.0000000000001071
26 Abell B, Glasziou P, Hoffmann T. Reporting and replicating trials of exercise-based cardiac rehabilitation: do we know what the researchers actually did?*Circ Cardiovasc Qual Outcomes* 2015;8:187-94. doi:10.1161/CIRCOUTCOMES.114.001381
27 Aveyard P, Johnson C, Fillingham S, Parsons A, Murphy M. Nortriptyline plus nicotine replacement versus placebo plus nicotine replacement for smoking cessation: pragmatic randomised controlled trial. *BMJ* 2008;336:1223-7. doi:10.1136/bmj.39545.852616.BE
28 Karanicolas PJ, Montori VM, Schünemann HJ, Guyatt GH. ACP Journal Club. "Pragmatic" clinical trials: from whose perspective?*Ann Intern Med* 2009;150:JC6-2, JC6-3. doi:10.7326/0003-4819-150-12-200906160-02002
29 Cahill K, Stevens S, Perera R, Lancaster T. Pharmacological interventions for smoking cessation: an overview and network meta-analysis. *Cochrane Database Syst Rev* 2013;2013:CD009329. doi:10.1002/14651858.CD009329.pub2
30 Sigal RJ, Kenny GP, Boulé NG, et al. Effects of aerobic training, resistance training, or both on glycemic control in type 2 diabetes: a randomized trial. *Ann Intern Med* 2007;147:357-69. doi:10.7326/0003-4819-147-6-200709180-00005
31 Church TS, Blair SN, Cocreham S, et al. Effects of aerobic and resistance training on hemoglobin A1c levels in patients with type 2 diabetes: a randomized controlled trial. *JAMA* 2010;304:2253-62. doi:10.1001/jama.2010.1710
32 Unverzagt S, Moldenhauer I, Nothacker M, et al. Immunotherapy for metastatic renal cell carcinoma. *Cochrane Database Syst Rev* 2017;5:CD011673. doi:10.1002/14651858.CD011673.pub2
33 Goldkuhle M, Guyatt GH, Kreuzberger N, et al. GRADE concept 4: rating the certainty of evidence when study interventions or comparators differ from PICO targets. *J Clin Epidemiol* 2023;159:40-8. doi:10.1016/j.jclinepi.2023.04.018
34 Peters AL, Ahmann AJ, Battelino T, et al. Diabetes Technology-Continuous Subcutaneous Insulin Infusion Therapy and Continuous Glucose Monitoring in Adults: An Endocrine Society Clinical Practice Guideline. *J Clin Endocrinol Metab* 2016;101:3922-37. doi:10.1210/jc.2016-2534
35 McCall AL, Lieb DC, Gianchandani R, et al. Management of Individuals With Diabetes at High Risk for Hypoglycemia: An Endocrine Society Clinical Practice Guideline. *J Clin Endocrinol Metab* 2023;108:529-62. doi:10.1210/clinem/dgac596
36 Korytkowski MT, Muniyappa R, Antinori-Lent K, et al. Management of Hyperglycemia in Hospitalized Adult Patients in Non-Critical Care Settings: An Endocrine Society Clinical Practice Guideline. *J Clin Endocrinol Metab* 2022;107:2101-28. doi:10.1210/clinem/dgac278
37 Mann H, Djulbegovic B. Comparator bias: why comparisons must address genuine uncertainties. *J R Soc Med* 2013;106:30-3. doi:10.1177/0141076812474779
38 Goldberg NH, Schneeweiss S, Kowal MK, Gagne JJ. Availability of comparative efficacy data at the time of drug approval in the United States. *JAMA* 2011;305:1786-9. doi:10.1001/jama.2011.539
39 Estellat C, Ravaud P. Lack of head-to-head trials and fair control arms: randomized controlled trials of biologic treatment for rheumatoid arthritis. *Arch Intern Med* 2012;172:237-44. doi:10.1001/archinternmed.2011.1209
40 Psaty BM, Weiss NS, Furberg CD. Recent trials in hypertension: compelling science or commercial speech?*JAMA* 2006;295:1704-6. doi:10.1001/jama.295.14.1704
41 Hunter RH, Joy CB, Kennedy E, Gilbody SM, Song F. Risperidone versus typical antipsychotic medication for schizophrenia. *Cochrane Database Syst Rev* 2003;(2):CD000440. doi:10.1002/14651858.CD000440
42 Safer DJ. Design and reporting modifications in industry-sponsored comparative psychopharmacology trials. *J Nerv Ment Dis* 2002;190:583-92. doi:10.1097/00005053-200209000-00002
43 Cohen JA, Barkhof F, Comi G, et al, TRANSFORMS Study Group. Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis. *N Engl J Med* 2010;362:402-15. doi:10.1056/NEJMoa0907839
44 Hauser SL, Bar-Or A, Comi G, et al, OPERA I and OPERA II Clinical Investigators. Ocrelizumab versus Interferon Beta-1a in Relapsing Multiple Sclerosis. *N Engl J Med* 2017;376:221-34. doi:10.1056/NEJMoa1601277
45 Cohen JA, Comi G, Selmaj KW, et al, RADIANCE Trial Investigators. Safety and efficacy of ozanimod versus interferon beta-1a in relapsing multiple sclerosis (RADIANCE): a multicentre, randomised, 24-month, phase 3 trial. *Lancet Neurol* 2019;18:1021-33. doi:10.1016/S1474-4422(19)30238-8
46 Kappos L, Wiendl H, Selmaj K, et al. Daclizumab HYP versus Interferon Beta-1a in Relapsing Multiple Sclerosis. *N Engl J Med* 2015;373:1418-28. doi:10.1056/NEJMoa1501481
47 Mohyuddin GR, Koehn K, Sborov D, et al. Quality of control groups in randomised trials of multiple myeloma enrolling in the USA: a systematic review. *Lancet Haematol* 2021;8:e299-304. doi:10.1016/S2352-3026(21)00024-7
48 Djulbegovic B, Lacevic M, Cantor A, et al. The uncertainty principle and industry-sponsored research. *Lancet* 2000;356:635-8. doi:10.1016/S0140-6736(00)02605-2
49 Zhang Y, Chen D, Cheng S, et al. Use of suboptimal control arms in randomized clinical trials of investigational cancer drugs in China, 2016-2021: An observational study. *PLoS Med* 2023;20:e1004319. doi:10.1371/journal.pmed.1004319
50 Ali DS, Mirza RD, Alsarraf F, et al. Systematic Review: Efficacy of Medical Therapy on Outcomes Important to Adult Patients with X-Linked Hypophosphatemia. *J Clin Endocrinol Metab* 2024;dgae890. doi:10.1210/clinem/dgae890
51 Singh AD, Mian A, Devasenapathy N, Guyatt G, Karthikeyan G. Percutaneous mitral commissurotomy versus surgical commissurotomy for rheumatic mitral stenosis: a systematic review and meta-analysis of randomised controlled trials. *Heart* 2020;106:1094-101. doi:10.1136/heartjnl-2019-315906
52 Walter SD, Sun X, Heels-Ansdell D, Guyatt G. Treatment effects on patient-important outcomes can be small, even with large effects on surrogate markers. *J Clin Epidemiol* 2012;65:940-5. doi:10.1016/j.jclinepi.2012.02.012