Core GRADE 5: Rating certainty of evidence – Assessing Indirectness

Gordon Guyatt, distinguished professor [1 2 3], Alfonso Iorio, professor [1 2], Hans De Beer, methodologist [4], Andrew Owen, professor [5], Thomas Agoritsas, associate professor [1 6 7], M Hassan Murad, professor [8], Ganesan Karthikeyan, professor, executive-director [9 10], Carlos Cuello, assistant professor [1], Manya Prasad, associate professor [11], Kevin Kim methodologist [1 12], Dalal S. Ali, clinical endocrinologist [13], Arnav Agarwal, methodologist [1 2 3], Lars G. Hemkens, professor [14 15 16], Liang Yao, assistant professor [17], Monica Hultcrantz, head of HTA Region Stockholm [18 19], Jamie Rylance, associate professor [20], Derek K. Chu, assistant professor [1 2], Per Olav Vandvik, professor [3 21 22], Benjamin Djulbegovic, professor, director [23], Reem A Mustafa, professor [1 24], Linan Zeng, professor [25], Prashanti Eachempati, adjunct Professor [3 26 27], Bram Rochwerg, associate professor [1 2], Kameshwar Prasad, professor emeritus [28 29], Victor Montori, professor [30 31], 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. Guide2Guidance, Lemelerberg 7, 3524 LC Utrecht, the Netherlands
5. Department of Pharmacology and Therapeutics, Centre of Excellence in Long-acting Therapeutics (CELT), University of Liverpool, Liverpool, United Kingdom.
6. MAGIC Evidence Ecosystem Foundation, Oslo, Norway
7. Division General Internal Medicine, University Hospitals of Geneva, Geneva, Switzerland
8. Evidence-based Practice Center, Mayo Clinic, Rochester, MN 55905, USA
9. Translational Health Science Technology Institute, Faridabad
10. All India Institute of Medical Sciences, New Delhi
11. Clinical Research and Epidemiology, Institute of Liver and Biliary Sciences, New Delhi, India
12. Population Health Research Institute, Hamilton, Canada
13. Divisions of Endocrinology and Metabolism, McMaster University, Hamilton, ON, Canada
14. Pragmatic Evidence Lab, Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland
15. Department of Clinical Research, University Hospital Basel and University of Basel, Basel, Switzerland
16. Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, USA
17. Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore
18. HTA Region Stockholm, Centre for Health Economics, Informatics and Health Care Research (CHIS), Stockholm Health Care Services, Sweden
19. Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden
20. Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, United Kingdom
21. Institute of Health and Society, University of Oslo Faculty of Medicine, Oslo, Norway
22. Department of Medicine, Lovisenberg Diakonale Hospital, Oslo, Norway
23. Division of Hematology/Oncology, Department of Medicine, Medical University of South Carolina 39 Sabin Street MSC 635 Charleston, SC 29425
24. Department of Medicine, University of Kansas Medical Center, Kansas City, MO, USA
25. Pharmacy Department/Evidence-based Pharmacy Centre/Children's Medicine Key Laboratory of Sichuan Province, West China Second University Hospital, Sichuan University; Sichuan University and Key Laboratory of Birth Defects and Related Disease of Women and Children, Ministry of Education; West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China
26. Peninsula Dental School, University of Plymouth, United Kingdom
27. Faculty of Dentistry, Manipal University College Malaysia
28. Department of Neurology, All India Institute of Medical Sciences, New Delhi, India

29. Dean (Research), 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

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

Take Home Messages

- GRADE distinguishes between two types of indirectness: indirect comparisons from network meta-analyses and indirectness related to PICO (Population, Intervention, Comparator, Outcome) elements. The latter is the focus of Core GRADE.
- Indirectness occurs when there is a mismatch between the PICO elements in the clinical question – the target PICO - and the 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 population mismatches include differences in age, changes in population over time, and differences in condition or disease between target PICO and study PICOs.
- Possible intervention and comparator mismatches include the intensity or duration of the intervention, and sub-optimal choice (most often of the comparator).
- Possible outcome mismatches include duration of follow-up and, most concerning, use of a surrogate outcome.

Curricular Objectives

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

- Understand two types of GRADE indirectness—indirect comparisons and indirectness related to patient/intervention/comparator/outcome (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);
- Identify and evaluate the extent of indirectness of the particular indirectness issues that arise for each of population, intervention and comparator, and outcomes.

Abstract

Guideline developers and health technology assessment (HTA) practitioners must carefully specify the patient/intervention/comparator/outcome (PICO) elements in their question of interest – 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.

Searches for direct evidence often yield evidence that does not completely match the target PICO. Indirectness observed in such situations is seldom serious and does not typically warrant rating down for indirectness. Serious concerns that do arise in searches of direct evidence include non-adherence to interventions; problematic comparators; studies in which patients allocated to the comparator receive the intervention; and studies that focus on surrogate rather than patient-important outcomes.

When direct evidence proves of low or very low certainty, systematic review authors preparing evidence to inform guidelines or HTA assessments may seek evidence from studies in which the PICO differs substantially from their target PICO. Such evidence may come from patient populations that differ in characteristics such as age or sex, or in their underlying condition, or from interventions, comparators or outcomes that differ from the target.

Whether examining studies emerging 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 and consider whether to rate down evidence certainty for indirectness.

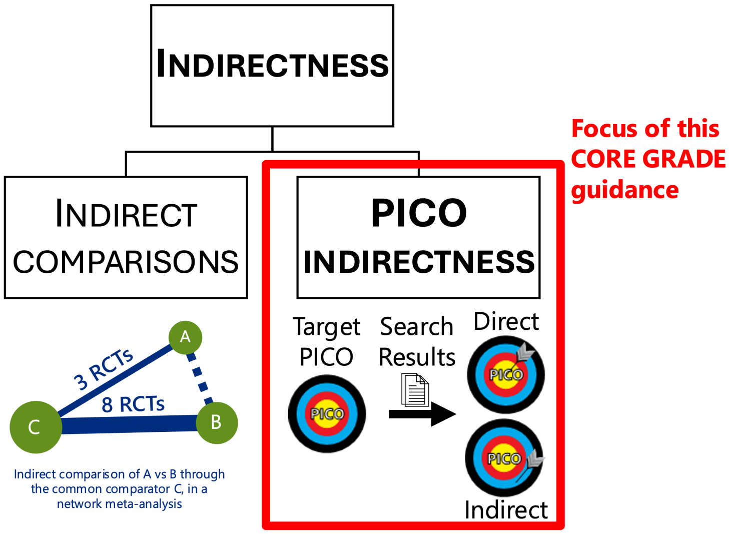
1. Introduction

This is the fifth article in a series of papers presenting Core GRADE, the essentials of applying GRADE to the conduct and evaluation of systematic reviews, clinical practice guidelines and health technology assessments (HTA) for questions of effects of interventions. The prior articles presented an overview of the GRADE process and discussed specific aspects of imprecision,[1] inconsistency,[2] and risk of bias.[3] This paper deals with the issue of what GRADE refers to as indirectness, and presents in sequence (a) the two types of GRADE indirectness – indirect comparisons, and indirectness related to patient/intervention/comparator/outcome (PICO) issues; (b) different importance of indirectness in systematic reviews, guidelines, and HTA; (c) distinguishing indirectness from inconsistency; (d) indirectness that arises in searching for direct evidence vs. relevant indirect evidence; (e) ensuring due attention to indirect evidence and (f) examples of the various sources of indirectness.

1.1 Two types of Indirectness – indirect comparisons

Previous GRADE guidance has used the term "indirectness" in two ways (Figure 1). In one, which we label "indirect comparisons", the interest is in the relative merits of intervention A versus intervention B but seeks evidence not from direct or head-to-head comparisons of A versus B but rather A versus C and B versus C comparisons. In this approach if, for instance, one finds that A does far better against C than does B, one infers superiority of A over B.

Figure 1. Two Types of Indirectness


INDIRECTNESS
INDIRECT COMPARISONS
PICO INDIRECTNESS
Focus of this CORE GRADE guidance
A
3 RCTs
8 RCTs
C
B
Indirect comparison of A vs B through the common comparator C, in a network meta-analysis
Target PICO
Search Results
Direct
PICO
PICO
PICO
Indirect

Over the last 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 to a single comparator. This article will therefore not deal further with indirect comparisons. Subsequently, we focus on the topic of this article: indirectness related to PICO issues.

1.2 Two types of Indirectness – indirectness related to PICO concerns

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

Note that 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; have anticipated high adherence with the intervention and found only low adherence; may have anticipated one standard of care in the comparator and observed another; or may have 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 the PICO question. Henceforth, when we use the term "indirectness", we will be referring to indirectness related to the target PICO rather than indirect comparisons.

2. Indirectness concerns are much more likely in guidelines and HTA assessments than in systematic reviews

When researchers conduct systematic reviews independently of HTA assessments and guidelines, they can establish eligibility criteria that closely fit their target PICO and restrict their inclusion criteria accordingly. Indirectness is as a result infrequently a major concern in such reviews.

HTA practitioners and guideline developers do not have that luxury. They choose, or are presented with, questions of sometimes urgent relevance to their target audiences. They must therefore identify and summarize the current best evidence to address those questions, even if that evidence represents a poor or limited match to their target PICO.

3. Indirectness versus inconsistency

When Core GRADE users have high suspicion that intervention effects are likely to differ between their target PICO and available evidence, they should be alert to the possibility of rating down for indirectness. If they have no reason to believe that relative effects differ in men versus women, for different drug doses, or for outcomes measured over 1 versus 3 years, they will remain untroubled in applying results to women when most evidence comes from men; to a higher dose when evidence comes from a lower dose; or to 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 indirectness concerns.

We have dealt with this central issue – do effects differ across subgroups of patients and interventions – in our overview of Core GRADE [4] and our article addressing inconsistency.[2] There, we have 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 regarding 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 the very elderly and younger people, low dose and high dose, or long 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.[2 5]

On the other hand, if Core GRADE users are interested in effects in the very elderly but all or almost all evidence comes from the younger people; interested in low dose but all or almost all evidence comes from high dose; or interested 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 individuals, the high dose, and the short follow-up to make inferences regarding their target

PICO. The extent to which relative effects will differ across such variables becomes a matter of mechanistic reasoning rather than empirical investigation and is thus less secure.

4. Indirectness in two scenarios

4.1 Indirectness encountered during the 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. In particular, patients may be older or younger, have a different ethnic background or have a different distribution of comorbidities from the target population.

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.

There are, however, particular situations in which serious indirectness exists in studies that prove eligible in a search for direct evidence. These include non-adherence to interventions, studies that focus on surrogate rather than patient-important outcomes, and problematic comparators. We will deal with these issues later in our discussion.

4.2 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 PICO. When Core GRADE users deliberately search for indirect evidence, they will inevitably confront the issue of indirectness.

5. Neglect of indirect evidence

Clinical practice guideline developers sometimes mistakenly conclude that no evidence exists regarding a PICO of interest. Very low quality evidence may, however, exist 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, repurposing interventions at the onset of the COVID-19 pandemic. The misguided enthusiasm for hydroxychloroquine and ivermectin highlights the limitations of such indirect evidence and thus the cautious inferences that it demands.

Indirect evidence may, even after considering rating down for indirectness, offer low or even moderate certainty. Even if it offers only very low certainty, that is preferable to making conclusions or decisions based on no evidence. Nevertheless, guideline developers sometimes neglect to consider indirect evidence. For example, a panel may conclude there is no evidence regarding a potential intervention in children. They will often, however, be thinking exclusively of direct evidence; they may be able to utilize indirect evidence from adults, as pediatricians would likely do in their clinical practice. In another example, early in the COVID-19 pandemic, no direct evidence for a number of interventions existed, but indirect evidence from related conditions (e.g. ARDS critically ill non-COVID patients) did exist, and provided support for guideline recommendations.[6]

Guideline developers not clear on the concept may use indirect evidence without explicitly labelling what they are doing. A study that evaluated guideline recommendations labeled as expert opinion found that most of these recommendations were in fact based on indirect evidence.[7]

Bearing in mind the possibility of indirect evidence, guideline developers and HTA 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.

6. Examples of Indirectness: Differences in population

6.1 Differences in age

Differences in age groups constitute a common indirectness issue in patients: elderly versus younger individuals or children versus adults. For example, in a guideline that addressed the management of pediatric pancreatitis, authors found very limited evidence for antibiotic use in children. They therefore conducted a search for evidence from adults, ultimately using the indirect evidence as the basis for their recommendations.[8]

6.2 Changes over time

Target patients may differ in many ways from patients enrolled in research studies. For example, the nature of the presenting patients may evolve over time, as occurred during the COVID-19 pandemic. Casirivimab and imdevimab given as a combination, and sotrovimab given alone, are monoclonal antibodies that bind to the SARS-CoV-2 spike protein, thus neutralizing the virus. Randomized control trials (RCTs) conducted in 2020 and 2021 demonstrated that both the combination of casirivimab and imdevimab, and sotrovimab reduced mortality in patients infected with the circulating virus at that time, motivating World Health Organization (WHO) recommendations for use of these agents.

However, changes in the sequence of the virus spike protein that occurred when omicron or its subsequent sub-lineages became the dominant variants resulted in markedly diminished in vitro neutralization activity.[9] The population in the target PICO had now changed from those infected with the viruses circulating earlier to those infected with the subsequently circulating variants. The panel had no direct evidence available – that is, no studies of the antibodies in the era of the new virus variants. Thus, the laboratory evidence of diminished antibody neutralization led the panel to conclude that the original RCTs now provided only very indirect evidence for the key outcomes, the antibodies were very unlikely to be effective in the new target population, and that strong recommendations against use of the antibodies were now warranted.[10]

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 therapies differed. Relapsing and remitting multiple sclerosis provide an example of this phenomenon.[11]

6.3 Differences in condition

On occasion, when direct evidence is unavailable or of low or very low certainty, authors of systematic reviews 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 with valvular heart disease requiring surgery.[12] 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 RCTs of warfarin versus placebo in other populations. They found an incidence rate ratio of bleeding 2.99, 95% CI 1.46 – 6.13 which, after rating down for indirectness of the population, they considered moderate certainty evidence of increased bleeding with the mechanical heart valves.

6.3.1 Indirect evidence regarding harms

In rare conditions, RCTs are typically small or very small. Estimates of intervention harms may therefore yield very wide confidence intervals 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 generally 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[13] and chronic urticaria[14] and allergen immunotherapy in asthma and allergic rhinitis to its use in atopic dermatitis.[15]

Core GRADE users have also applied the same principle to related conditions to improve the precision (i.e., narrow confidence intervals) 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. [16]

7. Examples of Indirectness: Differences in interventions

Interventions studied may differ from the target PICO in a number of ways including dose of a drug (higher or lower than the target), duration of administration (shorter or longer), route of administration (parenteral versus oral), or the skill level of providers of some interventions, such as educational, surgical, physical therapy, and psychosocial interventions. Another distressingly common source of indirectness for such interventions is that authors may not sufficiently describe the components of the interventions, which can make replication of these interventions impossible. For instance, description of cardiac rehabilitation details was so poorly reported in the literature that surveys of rehabilitation programs showed that what they implemented in practice differed substantially from what RCTs have demonstrated effective.[17] Inadequate description of the intervention constitutes a reason for rating down for indirectness.

7.1 Non-adherence

Another common way that trials of interventions differ from the target is patient non-adherence. Generally, patients and their health care 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, an RCT of nortriptyline as an adjunct to nicotine replacement for smoking cessation randomized 901 adult patients attending a smoking cessation service to nortriptyline or placebo.[18] They found that one year after "quit day", 11% in the nortriptyline group vs. 9% in the control group (relative risk 1.26, 95% CI 0.84 – 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 medications. Had there been close to 100% adherence, 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 the nortriptyline on smoking cessation in those who use the intervention.[19] A systematic review that included additional trials which also had adherence concerns narrowed the confidence interval (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).[20]

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 favor of nortriptyline. It is entirely plausible that, had adherence been very high, results would have demonstrated a benefit of nortriptyline in improving quit rates in smokers trying to end their addiction.

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 occur particularly often in HTA assessments.

When the target PICO focuses on those who adhere, indirectness due to non-adherence is often a major issue in trials of screening interventions. Consider an RCT of colonoscopy screening for colorectal cancer versus no such screening and a PICO of interest that specifies that patients all undergo the screening intervention. This is very plausibly the question of interest to patients: 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 program in which only some of the eligible population will be interested.

An RCT of colonoscopy allocated over 84,000 participants in Norway (highest participation over 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 approximately 2 in 1,000 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 and providing an estimate of the relative risk reduction that patients would experience if compliant showed a 45% relative risk reduction. With this estimate, the effect is still small but appreciably greater, approximately 6 per 1,000. The per protocol analysis provides a more direct estimate but with increased risk of bias.

Finally, it is possible that RCTs in which patients achieved high adherence may provide indirect evidence from a public health or funder's point of view. Studies of behavioral 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 high adherence unrealistic for clinical practice.[21 22] 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.

7.2 Trials that allow switching treatments

Oncology trials may have protocols that allow switching treatments when a patient fails the original intervention. For instance, consider the relative effects of two anti-cancer drugs, interferon-alpha and sunitinib in adults suffering from renal cancer. Systematic review authors encountered trials in which participants who experienced disease progression following interferon-alpha received sunitinib and other related target therapies. [23] How might this design bear on issues of indirectness?

The answer lies in considering the target PICO.[24] Core GRADE users whose PICO designates the comparison of sunitinib alone to interferon-alpha with the provision that patients who fail with the drug are offered sunitinib will find results directly applicable. 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 intervention arm participants that 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 have 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 there is substantial switching to sunitinib, and the arm that began with interferon-alpha proves similar to the sunitinib arm, the issue is in doubt: is the "rescue" sunitinib responsible for the similar results, or would they have been achieved with interferon-alpha alone? On the other hand, if sunitinib proves superior, that superiority would only have been greater had there been no switching. In the relevant systematic review sunitinib and other related target therapies proved superior to interferon-alpha (relative survival 1.3, 95% CI 1.13 – 1.51). Thus, indirectness does not compromise the conclusion regarding sunitinib's superiority to interferon-alpha, and authors have no need to rate down for indirectness.

7.3 Change of intervention technology

When the intervention is a device or a technology, its evolution over time can result in important indirectness that lowers certainty. For example, devices that help individuals care for their diabetes are constantly changing. Continuous glucose monitoring systems were approved by the Food and Drug Administration in the late 1990's and have quickly evolved with new sensor technology to lengthen their wear time from a few days to weeks to 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 vs. including studies of older systems and lowering certainty due to indirectness.[25-27]

8. Examples of Indirectness: Differences in comparators

Situations in which the comparator differs from the target PICO include differences in standard care in different jurisdictions; use of placebo when an active treatment is the clinically relevant comparator; inferior older alternatives rather than current optimal alternatives; and differences in dose or route of administration.[28] 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,[29] particularly as regulators require some form of placebo control in early phase 3 drug trials. For example, many RCTs of biologic disease-modifying drugs for rheumatoid arthritis patients did not use active comparators,[30] 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, suffer from indirectness.

The use of suboptimal comparators in industry trials is common, including the following examples. Large industry-sponsored trials evaluating newer antihypertensive drugs chose the beta-blocker atenolol as the comparator, despite prior evidence demonstrating inferiority of beta-blockers to a low-dose thiazide diuretic.[31]

Manufacturers of newer antipsychotic agents overestimated reduced toxicity advantages of their drugs by comparing them to inappropriately large doses of older alternatives.[32] Eight such trials used fixed doses of haloperidol of 20 mg per day, substantially above recommended doses.[33] A number of 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.[34-37]

A more recent example comes from RCTs in patients with multiple myeloma conducted in the United States in which enrolment occurred between 2010 and 2020.[38] The authors considered a control group regimen inferior if a previous RCT had shown an improved progression-free survival versus the control group before enrolment. Of 49 identified RCTs, seven (14%) began enrolling patients into inferior control groups after 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.[39]

Chinese investigators studying RCTs of cancer drugs authorized 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 RCTs adopted a suboptimal control arm.[40] 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 hypophosphatemia.[41] In particular, they wished to evaluate the impact of burosumab on pain and function, both against no specific treatment and against conventional therapy of phosphate salts and active Vitamin D. They identified an RCT of burosumab versus placebo that provided moderate to high certainty evidence for some of the key outcomes, but no study comparing the drug to 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. While not a satisfactory situation, the authors approach is the best possible under the circumstances.

9. Differences in outcomes

The impact of intervention versus comparators on outcomes may differ in how the outcomes are measured (for instance symptomatic versus asymptomatic vertebral fracture or symptomatic versus asymptomatic deep venous thrombosis), or the duration of follow-up (short-term versus long-term). They may also differ in whether the outcomes 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 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 stenosis faced with the choice of percutaneous versus surgical mitral commissurotomy. A key outcome for such patients is progression of heart failure symptoms due to failure of the procedure over the long term. Typically, 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 randomized trials comparing the two procedures addressed their relative merits for minimizing development or progression of heart failure symptoms. [42]

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 catheterization. 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, but also for indirectness of the outcome, resulting in very low certainty evidence as demonstrated in an adaptation of their summary of finding table (Table 1).

Table 1: Summary of findings table addressing long-term heart failure symptoms in RCTs of percutaneous versus mitral commissurotomy.

Outcome	Number of trials/patients	Result	Certainty of evidence
Heart failure symptoms as inferred from mitral valve area	6 RCTs, 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

While 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 by only one level for indirectness.[42] 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.

10. Conclusion

Limitations in the extent to which the PICO in the available studies differs from the target PICO – in GRADE called indirectness - represent a frequent reason for rating down certainty of evidence in the development of guidelines and HTA. When direct evidence is unavailable, or of low or very low certainty, Core GRADE users should 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 their target PICO, Core GRADE users must consider the likelihood that these differences will result in important variation in intervention effects, and if it is likely, should rate down once or - particularly with surrogate outcomes - twice for indirectness.

References

1. Guyatt G, Zeng L, Brignardello-Petersen R, et al. Core GRADE 2: Choosing the Target of Certainty Rating and Assessing Imprecision. *BMJ (in submission)* 2024
2. Guyatt G, Schandelmaier S, Brignardello-Petersen R, et al. Core GRADE 3: Rating Certainty of Evidence. Inconsistency. *BMJ (in submission)* 2024
3. Guyatt G. Core GRADE 4: Rating Certainty of Evidence. Risk of Bias. *BMJ (in submission)* 2024
4. Gyuatt G. Core GRADE 1: Overview of the Core GRADE Process *BMJ (in submission)* 2024
5. Schandelmaier S, Briel M, Varadhan R, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. *Cmaj* 2020;192(32):E901-e06. doi: 10.1503/cmaj.200077
6. 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(27):E756-E67.
7. 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(5):164-69. doi: 10.1136/ebmed-2017-110798 [published Online First: 20170918]
8. 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(1):159-76. doi: 10.1097/mpg.0000000000001715
9. 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(8):1117-18. doi: 10.1016/s1473-3099(22)00422-4 [published Online First: 20220628]
10. Agarwal A, Hunt B, Stegemann M, et al. Therapeutics and COVID-19: living guideline 2023 [Available from: https://app.magicapp.org/#/guideline/6989 accessed Sept 20 2024.
11. Zhang Y, Salter A, Wallström E, et al. Evolution of clinical trials in multiple sclerosis. *Ther Adv Neurol Disord* 2019;12:1756286419826547. doi: 10.1177/1756286419826547 [published Online First: 20190221]
12. 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(4):E450-e59. doi: 10.1503/cjs.001121 [published Online First: 20220712]
13. Zheng XL, Vesely SK, Cataland SR, et al. ISTH guidelines for the diagnosis of thrombotic thrombocytopenic purpura. *J Thromb Haemost* 2020;18(10):2486-95. doi: 10.1111/jth.15006 [published Online First: 20200911]
14. 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(4):437-44.e18. doi: 10.1016/j.anai.2024.06.003 [published Online First: 20240618]
15. 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(1):147-58. doi: 10.1016/j.jaci.2022.09.020 [published Online First: 20220930]
16. 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(4):e1071. doi: 10.1097/cce.0000000000001071 [published Online First: 20240401]
17. 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(2):187-94. doi: 10.1161/circoutcomes.114.001381 [published Online First: 20150303]

18. Aveyard P, Johnson C, Fillingham S, et al. Nortriptyline plus nicotine replacement versus placebo plus nicotine replacement for smoking cessation: pragmatic randomised controlled trial. *Bmj* 2008;336(7655):1223-7. doi: 10.1136/bmj.39545.852616.BE [published Online First: 20080427]
19. Karanicolas PJ, Montori VM, Schünemann HJ, et al. ACP Journal Club. "Pragmatic" clinical trials: from whose perspective? *Ann Intern Med* 2009;150(12):Jc6-2, jc6-3. doi: 10.7326/0003-4819-150-12-200906160-02002
20. Cahill K, Stevens S, Perera R, et al. Pharmacological interventions for smoking cessation: an overview and network meta-analysis. *Cochrane Database Syst Rev* 2013;2013(5):Cd009329. doi: 10.1002/14651858.CD009329.pub2 [published Online First: 20130531]
21. 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(6):357-69. doi: 10.7326/0003-4819-147-6-200709180-00005
22. 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(20):2253-62. doi: 10.1001/jama.2010.1710
23. Unverzagt S, Moldenhauer I, Nothacker M, et al. Immunotherapy for metastatic renal cell carcinoma. *Cochrane Database Syst Rev* 2017;5(5):Cd011673. doi: 10.1002/14651858.CD011673.pub2 [published Online First: 20170515]
24. 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;159:40-48. doi: 10.1016/j.jclinepi.2023.04.018 [published Online First: 20230503]
25. 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(11):3922-37. doi: 10.1210/jc.2016-2534 [published Online First: 20160902]
26. 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(3):529-62. doi: 10.1210/clinem/dgac596
27. 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(8):2101-28. doi: 10.1210/clinem/dgac278
28. Mann H, Djulbegovic B. Comparator bias: why comparisons must address genuine uncertainties. *J R Soc Med* 2013;106(1):30-3. doi: 10.1177/0141076812474779
29. Goldberg NH, Schneeweiss S, Kowal MK, et al. Availability of comparative efficacy data at the time of drug approval in the United States. *Jama* 2011;305(17):1786-9. doi: 10.1001/jama.2011.539
30. 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(3):237-44. doi: 10.1001/archinternmed.2011.1209
31. Psaty BM, Weiss NS, Furberg CD. Recent trials in hypertension: compelling science or commercial speech? *Jama* 2006;295(14):1704-6. doi: 10.1001/jama.295.14.1704
32. Hunter RH, Joy CB, Kennedy E, et al. Risperidone versus typical antipsychotic medication for schizophrenia. *Cochrane Database Syst Rev* 2003(2):Cd000440. doi: 10.1002/14651858.Cd000440
33. Safer DJ. Design and reporting modifications in industry-sponsored comparative psychopharmacology trials. *J Nerv Ment Dis* 2002;190(9):583-92. doi: 10.1097/00005053-200209000-00002
34. Cohen JA, Barkhof F, Comi G, et al. Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis. *N Engl J Med* 2010;362(5):402-15. doi: 10.1056/NEJMoa0907839 [published Online First: 20100120]

35. Hauser SL, Bar-Or A, Comi G, et al. Ocrelizumab versus Interferon Beta-1a in Relapsing Multiple Sclerosis. *N Engl J Med* 2017;376(3):221-34. doi: 10.1056/NEJMoa1601277 [published Online First: 20161221]

36. Cohen JA, Comi G, Selmaj KW, et al. 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(11):1021-33. doi: 10.1016/s1474-4422(19)30238-8 [published Online First: 20190903]

37. Kappos L, Wiendl H, Selmaj K, et al. Daclizumab HYP versus Interferon Beta-1a in Relapsing Multiple Sclerosis. *N Engl J Med* 2015;373(15):1418-28. doi: 10.1056/NEJMoa1501481

38. 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(4):e299-e304. doi: 10.1016/s2352-3026(21)00024-7

39. Djulbegovic B, Lacevic M, Cantor A, et al. The uncertainty principle and industry-sponsored research. *Lancet* 2000;356(9230):635-8. doi: 10.1016/s0140-6736(00)02605-2

40. 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(12):e1004319. doi: 10.1371/journal.pmed.1004319 [published Online First: 20231212]

41. Ali DS, Mirza RD. Systematic Review: Efficacy of Medical Therapy on Outcomes Important to Adult Patients with X-Linked Hypophosphatemia. . *Submitted to JCEM* 2024

42. Singh AD, Mian A, Devasenapathy N, et al. Percutaneous mitral commissurotomy versus surgical commissurotomy for rheumatic mitral stenosis: a systematic review and meta-analysis of randomised controlled trials. *Heart* 2020;106(14):1094-101. doi: 10.1136/heartjnl-2019-315906 [published Online First: 20200123]

Contributors: Gordon Guyatt, Victor Montori, Thomas Agoritsas, Monica Hultcrantz and Alfonso Iorio conceived and designed the Core GRADE series. Gordon Guyatt, Alfonso Iorio and Romina Brignardello-Petersen 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 the public.