Core GRADE 1: overview of the Core GRADE approach

Gordon Guyatt,[1,2,3,*] Thomas Agoritsas,[1,3,4,*] Romina Brignardello-Petersen,[1] Reem A Mustafa,[1,5] Jamie Rylance,[6] Farid Foroutan,[1,7] Manya Prasad,[8] Arnav Agarwal,[1,2,3] Hans De Beer,[9] M Hassan Murad,[10] Stefan Schandelmaier,[11,12,13] Alfonso Iorio,[1,2] Liang Yao,[14] Roman Jaeschke,[1,2] Per Olav Vandvik,[3,15] Linan Zeng,[16,17,18] Sameer Parpia,[1] Rohan D'Souza,[1,19] David Rind,[20,21] Derek K Chu,[1,2] Prashanti Eachempati,[3,22,23] Kameshwar Prasad,[24,25] Monica Hultcrantz,[26,27] Victor M Montori[28,29]

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:e081903
http://dx.doi.org/10.1136/bmj-2024-081903

Accepted: 14 March 2025

This first article in a seven part series presents an overview of the essential elements of the GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach that has proved extremely useful in systematic reviews, health technology assessment reports, and clinical practice guidelines. GRADE guidance has appeared in many articles dealing with both core issues and more specialised and complex guidance, and it has evolved over time. This series of articles presents GRADE essentials, Core GRADE, focusing on the core judgments necessary to summarise the comparative evidence about alternative care options and to make recommendations that apply to the care of individual patients. This article presents detailed guidance on formulating questions using the PICO (population, intervention, comparison, outcome) structure, and refining the question considering possible differences in relative and absolute effects across patient groups. The article then provides an overview of the remainder of the Core GRADE approach, including decisions about the certainty of the evidence and considerations in moving from evidence to guidance and recommendations.

SUMMARY POINTS

Core GRADE provides guidance for comparing a single intervention with a single comparator

Core GRADE takes primarily an individual patient rather than a population or public health perspective

Core GRADE deals with two fundamental processes: judging the certainty of evidence for each important outcome as presented in a summary of findings table, and moving from evidence to recommendations that have both a direction (for or against) and a strength (strong or conditional/weak)

The first step in the GRADE approach is to explicitly define the question using the PICO (population, intervention, comparison, outcome) format and then to refine the question considering the possibility of both subgroup differences in relative effect and large differences in baseline risk across patient groups

This is the first paper in a series describing Core GRADE (Grading of Recommendations Assessment, Development and Evaluation)—the fundamentals that authors of systematic reviews, health technology assessment reports, and clinical practice guidelines require to use the Core GRADE approach to rate certainty (quality, confidence) of evidence and grade strength of recommendations. Although the series focuses on evidence comparing a single intervention with a single comparator and takes an individual patient's perspective, the underlying principles also apply to evidence and decisons from a population or public health perspective.

Figure 1 depicts each of the key steps in using Core GRADE to create clinical practice guidelines or health technology assessment reports: summarising the evidence, rating its certainty, and moving from evidence to recommendations. We then provide details of the first steps in this process—formulating and refining the clinical question. Finally, we present an overview of the subsequent steps: rating the certainty of evidence and preparing tables with a summary of the findings. Subsequent papers in this series will address each of these key steps in detail. Supplementary Appendix 1 provides plain language definitions of key GRADE terms.

The information in this article will enable Core GRADE users to understand the scope and aim of this new series; describe Core GRADE's approach to rating certainty of evidence and summarising the evidence in systematic reviews, clinical practice guidelines, and health technology assessment reports; frame questions using the population, intervention, comparison, outcome (PICO) format, with consideration of possible differences across subgroups of patients and the possible need to seek indirect evidence; and describe Core GRADE's approach to moving from evidence to recommendations.

The GRADE approach

Core GRADE applies to the design, conduct, and analysis of systematic reviews as well as to the use

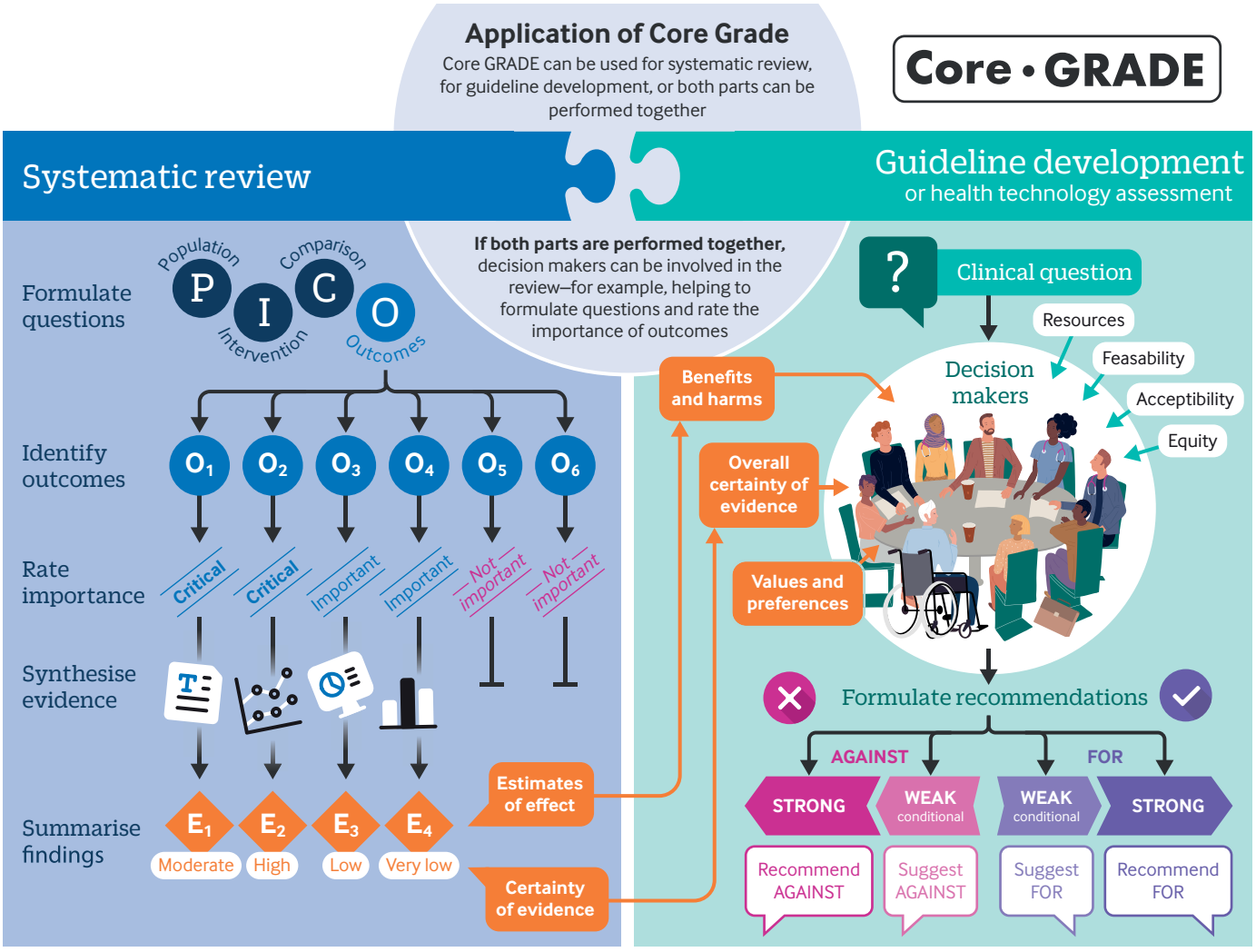

Fig 1 | Schematic overview of Core GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach

of their results in moving from evidence to decisions in clinical practice guidelines or health technology assessment.[1] In figure 1, the left panel addresses the systematic review process, including the definition of the review question formulated using the PICO format with specification of the relative importance of each of the components, followed by the collation and summarisation of the evidence, including ratings of certainty (quality) of the evidence for each outcome.

The right panel depicts the decision making process involved in the development of clinical practice guidelines and health technology assessment reports. That process involves a guideline panel or decision making group that considers key issues of magnitude of benefit, harms and burdens, certainty of evidence, and patient or public values and preferences. Decision makers might also consider costs, feasibility, acceptability, and equity in coming to recommendations either for or against an intervention and further specifying recommendations as strong or conditional (weak).

When we refer to recommendations we mean not only the highly structured guidance GRADE suggests for clinical practice guidelines but also the less formally structured guidance that health technology assessment reports typically provide for their target decision makers. Although we focus our discussion on the structure of the guideline development process, it is also relevant to health technology assessment and, excepting the process of moving from evidence to recommendations, to systematic reviews.

Formulating the question

All applications of GRADE, whether for systematic reviews, health technology assessment reports, or clinical practice guidelines, begin with the identification of a clinical question using a structured process to specify the relevant population, interventions, comparisons, and outcomes of potential importance to patients—the target PICO.

Optimal use of GRADE involves careful thought about the formulation of each element of the PICO

structure, being as specific as possible and avoiding ambiguity.

Population, intervention, and comparison

In the past, GRADE users have often lacked precision in how they specify their target PICO. The careful, explicit, and transparent specification of the target PICO represents a key aspect of Core GRADE. In particular, Core GRADE users must be as precise as possible when defining the target population.

Including a GRADE expert explicitly charged with helping systematic review teams and guideline panels to formulate their PICOs has proved helpful in achieving the desired precision. The GRADE expert should work with a steering group that translates broad questions generated by the panel into detailed structured questions. The steering group will subsequently help the panel consider differences between its target population and patients enrolled when the Core GRADE approach considers issues of indirectness. The fifth paper in this series addresses these issues.

Regarding choice of the population, the methods expert group of a World Health Organization (WHO) panel (with the function of a steering group in other settings) addressing the management of sepsis confronted problems in arriving at a precise definition—problems the GRADE expert on the group highlighted. Core GRADE users will often encounter studies with varied definitions of their population. In the current example, the guideline developers came across several definitions of sepsis, including a revised consensus definition that differed from that used in individual trials, which also differed from one another. To resolve the problem of such discrepancies, the panel used the most recent consensus definition of sepsis, recognising that the greater the extent to which eligibility criteria in the trials differed from the consensus, the greater the possibility that the evidence included in the review would not reflect the target population (and thus, in GRADE terminology, would introduce indirectness).

Supplementary appendix 2 presents another example of the need for careful specification in defining the PICO and highlights how authors must think ahead to the needs of their target audience when formulating the research question.[2]

Rating the importance of outcomes

Core GRADE users will initially strive to identify all outcomes that are potentially important to patients and subsequently consider those of sufficient importance. Guideline panels in particular may undertake a structured process to rate the importance of potentially relevant outcomes. In doing so, they will adopt the perspective of the population to which the guideline applies and may use a 9 point scale, with ratings of 7 to 9 representing critical outcomes, 4 to 6 representing important but not critical outcomes, and 1 to 3 representing unimportant outcomes.[3 4] Ideally, they would make this rating informed by a systematic review of patient values and preferences, and, if resources permit, a focus group of relevant patients.[5]

In the hypothetical guideline depicted in figure 1, panelists identified six outcomes of potential relevance, of which they judged two as critically important, two as important, and two as not important. They would give no further attention to the two outcomes considered not important while continuing to consider the other four.

Outcomes of importance can differ across the stages of an illness. For example, an Infectious Diseases Society of America guideline panel addressing patients admitted to hospital with covid-19 considered the most important outcomes (those GRADE labels as critical) as mortality, need for invasive mechanical ventilation, and duration of hospital admission.[6] For outpatients with covid-19, the panel considered hospital admission as a critical outcome. For patients receiving pre-exposure or postexposure prophylaxis, critical outcomes included measures of symptomatic covid-19.

Another crucial problem in considering outcomes is the relevant time frame: days, months, or years, and, whatever chosen, how many? In the WHO sepsis guideline, for all the wide variety of interventions under consideration (eg, corticosteroids, inotropes, different starting times for antibiotics) and outcomes (eg, mortality, adverse effects of treatment, functional recovery), the panel considered several alternatives: during hospital admission, the first month after hospital admission, or a considerably longer timeframe (eg, six months). The panel decided on the 28 days after hospital admission as the primary duration of interest and six months as a secondary duration.

Because the lowest certainty of evidence of the critical outcomes informs overall certainty of evidence, guideline panels—and the systematic reviews created to inform a specific guideline—need to distinguish between critical outcomes and important but not critical outcomes. Often, however, systematic review authors perform reviews independently of practice guidelines or health technology assessments. Although most problems faced by these independent reviews do not differ from reviews specifically conducted for health technology assessments or guidelines, one that does concerns differentiating between critical and important but not critical outcomes. Because systematic review authors do not need to make definitive judgments regarding the balance of benefits, harms, and burdens, they do not need to judge the relative importance of the outcomes they have chosen. In particular, they do not need to distinguish between critical outcomes and important but not critical outcomes.

Refining the question: possibly different effects across population, intervention, or outcome definitions

Core GRADE users will further refine the clinical question, considering the possibility that treatment effects differ across population, intervention, or outcome characteristics. In the past, GRADE users have often shown confusion between relative and absolute effects, and we will here clarify this problem in some detail. Furthermore, systematic review authors and

guideline panels have generated an excessive number of hypotheses and have failed to specify the direction of effect. In addition, they have often failed to recognise that differences in relative effect across subgroups of patients are unusual and have been ready to judge the credibility of subgroup effects as high when they are not.[7 8]

For binary (yes/no) outcomes such as death and hospital admission, treatment effects may differ as a result of differences in relative effects (eg, no effect—that is, relative risk of 1.0—in one patient group, relative risk of 0.5 in another), or differences in prognosis in the comparator group, which we often refer to as baseline risk. For example, the relative risk stays the same, but the comparator group event rate—the baseline risk—is 5% in one group of patients and 20% in another.

For differences in relative effects, authors refer to factors that could influence the magnitude of effect using interchangeable terms such as subgroup effect, effect modification, or interaction. For continuous outcomes such as pain or length of hospital stay, typically presented using an absolute difference scale, effects may also differ across groups. For example, in a PICO addressing pain using a 10 cm visual analogue scale, a difference between intervention and control of 3 cm in one patient group and 1 cm in another.

Patient characteristics that may modify either relative effect or baseline risk include age, severity of illness, patient circumstances such as socioeconomic status, and comorbidity. Intervention characteristics such as dose or duration of administration as well as outcome characteristics such as alternative outcome definitions or varying duration of follow-up can also influence treatment effects.

As they refine their question, GRADE users consider both relative effects (relative risks, odds ratios, or, for time to event analyses, hazard ratios) and absolute effects (risk differences). Relative treatment effects are, however, most often similar across subgroups—true subgroup relative effects are uncommon.[9-13] Core GRADE users should therefore be parsimonious in postulating subgroup relative effects and specify a direction for each postulated effect. For example, authors should not only identify that effects may differ in patients according to their age but also, for instance, that the intervention effects will be larger in elderly people than in younger people. Large numbers of hypothesised subgroup effects and failure to specify their direction will undermine the credibility of any apparent subgroup effect. A later article in this series addressing inconsistency in results across studies returns to the problem of credibility of subgroup effects.

Although relative effects are usually similar across patient subgroups, absolute effects of interventions will differ noticeably when subgroups of untreated patients have different probabilities (different baseline risks) of the outcomes of interest. For example, a treatment associated with a consistent relative risk reduction of 50% might reduce mortality in young people without comorbidity from 2% to 1% (absolute difference 1%) and in old people with comorbidity from 40% to 20% (absolute difference 20%). Such differences, although perhaps not often of this magnitude, are common. This is the reason that absolute and not relative effects are important to patients: young people in this example might be reluctant to use an intervention with substantial harms, whereas old people will likely be more inclined to do so. Supplementary appendix 3 presents an exercise in distinguishing between prognosis (baseline risk) and relative risk as sources of difference in absolute effects across groups.

Ideally, panels and review teams will specify a small number (≤3) of relative subgroup effects including a postulated direction. Core GRADE users may also anticipate different recommendations for groups at different baseline (absolute) risks and this might result in different recommendations—for example, recommendations in favour of an intervention for patients at higher risk and against for patients at lower risk.

In an example adapted from WHO's living guideline for covid-19 therapeutics,[14] when making

Table 1 | Formulating the PICO question for using nirmatrelvir-ritonavir in patients with non-severe covid-19, including differences in both baseline risks and relative risks between patient groups. Adapted from World Health Organization[14]

PICO and group specification	
Population	Patients with non-severe confirmed covid-19 (according to WHO severity definitions)
Intervention	Nirmatrelvir-ritonavir+standard care
Comparison	Standard care
Outcome	Hospital admission, mortality, mechanical ventilation, adverse effects leading to drug discontinuation, time to symptom resolution
Groups with possible varying baseline risk and thus different absolute effects (prognostic factors expected to lead to differences in baseline risk)	Likely determinants of baseline risk: Larger baseline risk and therefore larger beneficial effects of the intervention in immunosuppressed patients or those with chronic disease Larger baseline risk in those with negative serology results Larger baseline risk in unvaccinated people
Groups with possible varying relative risk (differences in effect in different patient groups)	Candidate effect modifiers: Larger beneficial relative effects of the intervention in those with shorter time from symptom onset Larger beneficial effects in those with negative serology results Larger beneficial effects in unvaccinated people

PICO=population, intervention, comparison, and outcome.

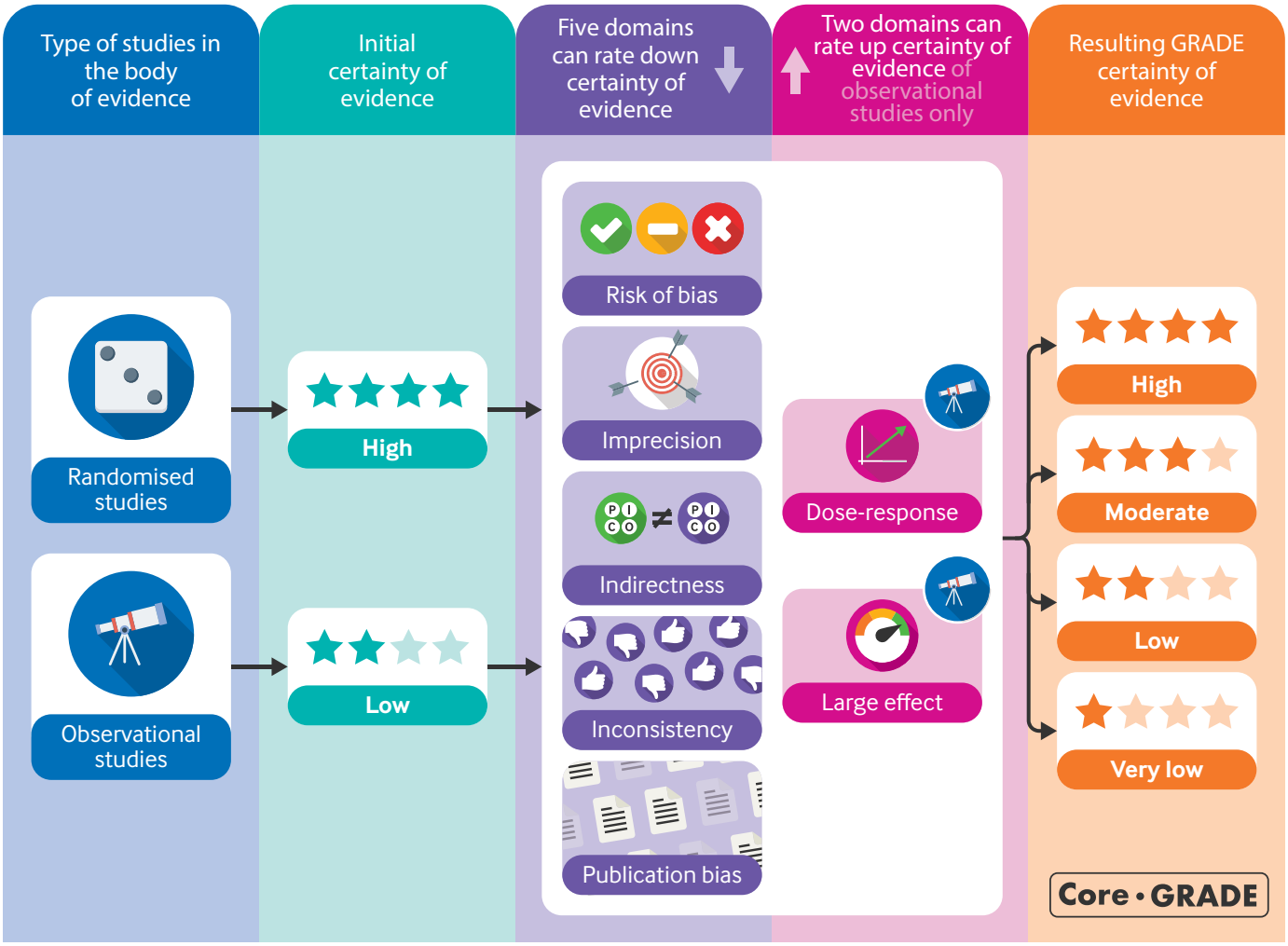

Fig 2 | Core GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach to rating certainty of evidence in intervention effects

recommendations for nirmatrelvir-ritonavir in patients with non-severe covid-19, the guideline panel postulated both possible absolute effects and possible relative differences between groups (table 1). For each postulated effect, the WHO panel made the direction of effect explicit, hypothesising for instance that the effect of treatment would diminish if the administration of the intervention were delayed.

The systematic review team reported no credible relative subgroup effects.[15] They did, however, identify substantial differences in effect between patients with different baseline risks. In particular, they found that despite patients with a high and low baseline risk having the same relative effect, nirmatrelvir-ritonavir was likely to effect important reductions in both mortality and hospital admissions in patients at high risk; in patients with a low baseline risk, benefits were likely to be trivial. More specifically, the anticipated reduction in hospital admissions for patients at low risk was about 4 per 1000 (0.4%) and for those at high risk was 50 per 1000 (5%). When considering the results of the systematic review showing the large differences in effect between patients at low and high risk, the guideline panel issued a strong recommendation in favour of nirmatrelvir-ritonavir for patients at high risk and a weak recommendation against its use for patients at low risk.

Possible need for indirect evidence

A final problem systematic review, health technology assessment, and clinical practice guidelines must consider is the possible need to seek indirect evidence—that is, evidence from studies with a PICO that does not completely correspond to the target PICO. This will be required when teams anticipate direct evidence may be limited and the highest certainty evidence may be indirect. For example, early in the covid-19 pandemic, before evidence was available from randomised controlled trials, guideline developers addressing the use of steroids for those with severe covid-19 considered evidence from randomised controlled trials showing benefit in patients with acute respiratory

distress syndrome whose clinical presentation is similar to that of critically ill patients with covid-19.[16] The panel decided that although the evidence was limited by indirectness, it warranted a conditional recommendation for use of steroids in patients with severe covid-19.[17]

Having decided on their PICO question and any subgroup hypotheses (based on either different baseline risks or different modifiers of relative effects), systematic review teams will, as depicted in figure 1, proceed to summarise the effects of the interventions of interest on the outcomes of interest in the patients of interest and then judge the associated certainty of evidence.

GRADE's rating of certainty of evidence

GRADE's certainty rating represents confidence that the true effect lies on one side of a chosen threshold (such as an important difference) or in a particular range (such as a small effect).[18] Core GRADE presents a four category system of rating certainty as high, moderate, low, and very low. Although Core GRADE certainty ratings rely on evaluation of individual studies, they refer to the entire body of evidence addressing a particular outcome rather than those individual studies.

Figure 2 summarises the Core GRADE approach to rating certainty of evidence for intervention effects. If the evidence comes from randomised controlled trials, ratings begin as high certainty. In contrast, a body of evidence from non-randomised studies of interventions (eg, cohort and case-control studies) begins as low certainty. Certainty in the evidence from both randomised controlled trials and non-randomised studies decreases when limitations are identified in any one of five domains: imprecision, inconsistency, risk of bias, indirectness, and publication bias. Core GRADE users can rate up certainty in non-randomised studies (but not randomised controlled trials) for large effects and for evidence of a dose-response gradient. Previous GRADE guidance that included the possibility of rating up as a result of predictable direction of plausible confounding has proved too difficult to apply and too rarely applicable to be part of Core GRADE.

We characterise limitations in each of these domains involved in rating down certainty as not serious; serious; very serious; or, rarely, extremely serious. The loss of certainty will result in rating down once for a particular domain (for example, from high to moderate certainty) if concerns are serious, and twice for a domain (for example, from high to low) if concerns are very serious.

Although GRADE has divided certainty of evidence into four categories, certainty is actually a continuum. As a result, one may, when rating is near a cut-off point between categories, have disagreements about certainty when judgments are in fact similar (fig 3).

In terms of ratings near a cut-off point, the same phenomenon can occur when deciding whether to rate down in any one domain. Ratings may be near the cut-off point between no serious and serious concerns (potentially mandating rating down one level) or between serious and very serious (potentially rating down two levels) (fig 4).

To illustrate the potential problem, consider the following scenario. While summarising certainty of evidence for randomised controlled trials, the rating is near the threshold between no serious and serious limitations for three domains, with no problems in the other two domains. Erring on the side of not rating down, one might not rate down for any of the three and emerge with high certainty evidence. Erring on the side of rating down, one might rate down for all three and emerge with very low certainty evidence.

In this situation, one might consider the magnitude of the problems in the three domains, conclude that the certainty of evidence is near the threshold between moderate and low certainty, and ultimately decide to either rate down once and conclude moderate certainty evidence or rate down twice and conclude low certainty evidence. Such a situation illustrates the necessity for, after considering each of the individual domains, stepping back and taking an overall view of the threats to certainty of evidence.

For example, in the WHO's living guideline for covid-19 therapeutics[14] recommendation about lopinavir-ritonavir versus standard care for mortality and mechanical ventilation, the review team identified problems with both inconsistency and imprecision for mortality and mechanical ventilation outcomes. In both cases, however, the problems were near the threshold between not serious and serious (fig 4) and thus were not so serious as to warrant rating down for both domains, with consequent low certainty of

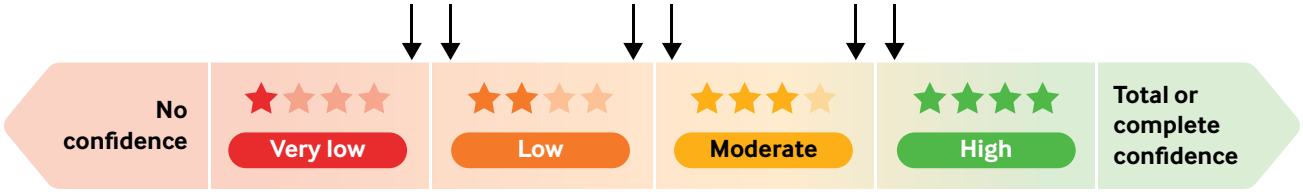

Fig 3 | Certainty of evidence is a continuum that GRADE (Grading of Recommendations Assessment, Development and Evaluation) divides into four categories. Making judgments about rating down certainty when near a cut-off point (arrows) can lead to differences in judgments when certainty is similar

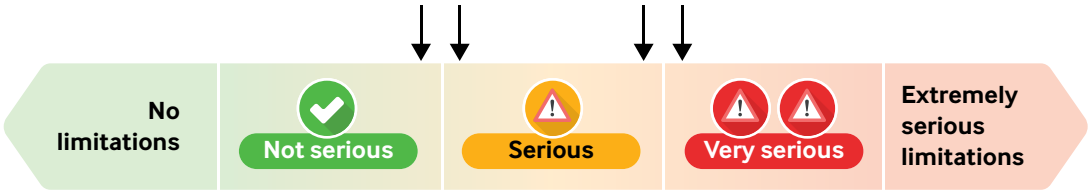

Fig 4 | Each factor for rating down or rating up certainty of evidence in GRADE (Grading of Recommendations Assessment, Development and Evaluation) reflects a continuum. Arrows represent choices near the cut-off points and thus represent apparent disagreement but true agreement

evidence (ie, rating down from high to low). Thus, the overall view for mortality and mechanical ventilation was that the evidence was of moderate certainty. The panel noted that concerns in both inconsistency and imprecision domains led to the rating down from high to moderate certainty evidence. These thoughtful judgments reflect the strength of Core GRADE: the facilitation of a thoughtful and deliberative assessment of evidence within a sound, carefully considered, transparent structure that allows for flexibility.

Completing the GRADE process: evidence summary and moving from evidence to recommendations

The review team will summarise the evidence in relative and absolute estimates of effect for each patient important outcome, including rating certainty of evidence as high, moderate, low, or very low, using a GRADE summary of findings table. Such tables may take a variety of formats, but any format should include a row for each outcome, and each row should include:

- key details of outcome measurement: the duration of time to which estimates apply and for continuous outcomes the scale, what higher and lower scores indicate, and, if available, a minimal important difference;
- the number of studies and number of patients contributing to the estimate of effect on that outcome;
- relative effects on binary outcomes presented as odds ratios, relative risks, and associated confidence intervals, time-to-event outcomes presented as hazard ratios and associated confidence intervals;
- for absolute effects, for binary outcomes: the baseline risk (event rate) in the control group, the event rate in the intervention group, the difference between these event rates and the confidence interval around that difference; for time to event outcomes: the median survival estimates or the event rate, or both, at a prespecified time during follow-up in each of the groups, and the difference between this estimates or rates with the confidence interval around that difference; and for continuous outcomes: the mean in each group and the difference between groups (typically a mean difference) and an associated confidence interval;
- the certainty of the evidence, with an indication of reasons for rating down if evidence is rated down from high certainty; and
- a plain language summary that captures the essential conclusion about the results, including the associated certainty.

Table 2 presents one way to structure summary of findings tables—our preferred format and one that WHO has been using in its guidelines. This example is adapted from a WHO guideline addressing nirmatrelvir-ritonavir versus standard care for patients with non-severe covid-19.[14 15] In a later article in this series we will provide more details on how to produce summary of findings tables.

For systematic reviews not linked to specific guidelines or health technology assessment reports, the GRADE process ends here. For clinical practice guidelines, the panel then considers the summary findings table informing its recommendation. An initial choice is to clearly differentiate the critical outcomes from the important but not critical outcomes. The lowest certainty of evidence of the critical outcomes will constitute the overall certainty of evidence for each recommendation.

To move from evidence to recommendations, the panel will consider the relative importance of the benefit, harm, and burden outcomes summarised in the summary of findings table. In doing so, they will consider not their own view of relative importance but rather the relative importance of the outcomes to the patients or people to whom the guideline applies. Core GRADE users make judgments about the relative importance of outcomes and how, as a result of the importance, patients will trade-off the benefits versus harms and burdens as values and preferences. Core GRADE mandates that the panel provide an explicit statement of the inferred values and preferences of the relevant population.

The panel will then address the balance of anticipated desirable and undesirable effects of the intervention and comparator, taking into account both underlying values and preferences and the certainty of evidence for each outcome. As depicted in figure 1, the panel will, if relevant to its particular question,

Table 2 | Summary of findings for nirmatrelvir-ritonavir in patients with non-severe covid-19. Adapted from World Health Organization[14 15]

Outcome	Study results and measurements	Absolute effect estimates		Certainty (quality) of the evidence	Plain language
		Nirmatrelvir-ritonavir	Standard care		
Mortality (low risk) over duration of illness	Odds ratio 0.04 (95% CI 0.00 to 0.67) Based on data from 2224 participants in one study	0.02 per 1000	0.5 per 1000 0.48 fewer per 1000 (95% CI 0.5 fewer to 0.16 more)	High	Nirmatrelvir-ritonavir does not result in an important reduction on mortality in low risk individuals
Mortality (high risk) over duration of illness	Odds ratio 0.04 (95% CI 0.00 to 0.67) Based on data from 2224 participants in one study	0 per 1000	6 per 1000 6 fewer per 1000 (95% CI 6 fewer to 2 fewer)	Moderate Owing to serious imprecision as CI crosses threshold of an important benefit	Nirmatrelvir-ritonavir probably reduces mortality in high risk individuals
Hospital admission (low risk) over duration of illness	Odds ratio 0.15 (95% CI 0.06 to 0.36) Based on data from 3078 participants in two studies	0.7 per 1000	5 per 1000 4.3 fewer per 1000 (95% CI 4.7 fewer to 3.2 fewer)	High (upper boundary of CI below threshold of importance)	Nirmatrelvir-ritonavir does not result in an important reduction in hospital admission in low risk individuals
Hospital admission (high risk) over duration of illness	Odds ratio 0.15 (95% CI 0.06 to 0.36) Based on data from 3078 participants in two studies	9 per 1000	60 per 1000 51 fewer per 1000 (95% CI 56 fewer to 38 fewer)	High	Nirmatrelvir-ritonavir reduces hospital admission in high risk individuals
Time to symptom resolution (days)	Lower better Based on data from 662 participants in one study	Median 9.0	Median 9.0 Difference in medians: 0 fewer (95% CI 3.0 fewer to 4.4 more)	Low Due to very serious imprecision	Nirmatrelvir-ritonavir may not reduce time to symptom resolution

CI=confidence interval.

consider cost, feasibility, acceptability, and equity, arriving at a direction of recommendation in favour of the intervention or comparator, and a strength of recommendation as strong or conditional (weak).

Limitations of the Core GRADE approach

We see only one serious limitation in the Core GRADE approach, and that is the subjectivity involved in virtually every important decision in rating certainty of evidence and moving from evidence to recommendations. This need for subjective judgments would, however, be inevitable for any useful system considering certainty of evidence or providing an approach to move from evidence to recommendations. The alternative—a mechanical approach offering never-to-be-broken rules—would result in many poor judgments and prove unfeasible and unacceptable.

Conclusion

For systematic reviews, health technology assessments, and clinical practice guidelines, the Core GRADE approach begins with formulating the relevant clinical questions using the PICO format. At the same time Core GRADE users will consider possible relative and absolute subgroup effects and provide a priori hypotheses about direction of effect for a small number of potentially credible effect modifiers. Reviewers summarise the evidence, including estimates of magnitude of relative and absolute effects and the certainty of the evidence, and present their findings in summary of findings tables. They may do so in the context of supporting a guideline, supporting other decision makers, or independent of either. In the decision making context, decision makers use review results to address core issues of benefits, harms and burdens, certainty of the evidence, and patient values and preferences. They do so whether evidence is high or low certainty, always responding to clinicians' and patients' need for guidance. Depending on the context, they might in addition consider secondary criteria of cost, acceptability, feasibility, and equity in deciding on recommending for or against an intervention versus a comparator, and the strength of the recommendation as strong or conditional.

AUTHOR AFFILIATIONS

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
[2]Department of Medicine, McMaster University, Hamilton, ON, L8S 4L8, Canada
[3]MAGIC Evidence Ecosystem Foundation, Oslo, Norway
[4]Division General Internal Medicine, University Hospitals of Geneva, University of Geneva, Geneva, Switzerland
[5]Department of Medicine, University of Kansas Medical Center, Kansas City, MO, USA
[6]Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK
[7]Ted Rogers Centre for Heart Research, University of Toronto, Toronto, ON, Canada
[8]Clinical Research and Epidemiology, Institute of Liver and Biliary Sciences, New Delhi, India
[9]Guide2Guidance, Utrecht, Netherlands
[10]Evidence-based Practice Center, Mayo Clinic, Rochester, MN, USA
[11]Division of Clinical Epidemiology, University Hospital and University of Basel, Basel, Switzerland
[12]School of Public Health, University College Cork, Cork, Ireland
[13]MTA–PTE Lendület "Momentum" Evidence in Medicine Research Group, Medical School, University of Pécs, Pécs, Hungary
[14]Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore
[15]Institute of Health and Society, University of Oslo Faculty of Medicine, Oslo, Norway
[16]Pharmacy Department/Evidence-based Pharmacy Centre/Children's Medicine Key Laboratory of Sichuan Province, West China Second University Hospital, Sichuan University, Chengdu, China
[17]Sichuan University and Key Laboratory of Birth Defects and Related Disease of Women and Children, Ministry of Education, Chengdu, China
[18]West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China
[19]Department of Obstetrics and Gynaecology, McMaster University, Hamilton, ON, Canada
[20]Institute for Clinical and Economic Review, Boston, MA, USA
[21]Harvard Medical School, Beth Israel Deaconess Medical Centre, Boston, MA, USA

[22]Peninsula Dental School, University of Plymouth, Plymouth, UK
[23]Faculty of Dentistry, Manipal University College Malaysia, Melaka, Malaysia
[24]Department of Neurology, All India Institute of Medical Sciences, New Delhi, India
[25]Fortis CSR Foundation, New Delhi, India
[26]HTA Region Stockholm, Centre for Health Economics, Informatics and Health Care Research (CHIS), Stockholm Health Care Services, Stockholm, Sweden
[27]Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden
[28]Division of Endocrinology, Department of Medicine, Mayo Clinic, Rochester, MN, USA
[29]Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN, USA
*Both authors contributed equally (joint first authors).

Contributors: GG and TA contributed equally and are joint first authors. GG, VMM, TA, MH, and AI conceived and designed the Core GRADE series. GG, TA, VMM, MH, RB-P, and FF drafted the manuscript. All authors critically revised the manuscript across several iterations for important intellectual content and gave final approval for the article. GG and TA are the guarantors. 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 Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011;64:383-94. doi:10.1016/j.jclinepi.2010.04.026.
2 Carson JL, Kleinman S. Indications and hemoglobin thresholds for RBC transfusion in adults; 2023. https://www.uptodate.com/contents/indications-and-hemoglobin-thresholds-for-rbc-transfusion-in-adults?sectionName=Acute%20MI&topicRef=15664&anchor=H812290&source=see_link#. [Accessed 21 Jan 2024.]
3 Guyatt G, Montori V, Devereaux PJ, Schünemann H, Bhandari M. Patients at the center: in our practice, and in our use of language. *ACP J Club* 2004;140:A11-2. doi:10.7326/ACPJC-2004-140-1-A11
4 Jaeschke R, Douketis J, Nowaczyk M, Guyatt G. Human-important outcomes and evidence-based medicine during the coronavirus disease 2019 pandemic. *Pol Arch Intern Med* 2020;130:714-5. doi:10.20452/pamw.15514.
5 Goodman SM, Miller AS, Turgunbaev M, et al. Clinical Practice Guidelines: Incorporating Input From a Patient Panel. *Arthritis Care Res (Hoboken)* 2017;69:1125-30. doi:10.1002/acr.23275.
6 Bhimraj A, Morgan RL, Shumaker AH, et al. Infectious Diseases Society of America Guidelines on the Treatment and Management of Patients with COVID-19. *Clin Infect Dis* 2020;ciaa478. doi:10.1093/cid/ciaa478.
7 Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ* 2012;344:e1553. doi:10.1136/bmj.e1553.
8 Schandelmaier S, Guyatt G. Same Old Challenges in Subgroup Analysis-Should We Do More About Methods Implementation? *JAMA Netw Open* 2024;7:e243339. doi:10.1001/jamanetworkopen.2024.3339.
9 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.
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 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.
12 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.
13 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.
14 World Health Organization. Therapeutics and COVID-19: living guideline, 13 January 2023: WHO; 2023. https://iris.who.int/handle/10665/365584
15 Siemieniuk RA, Bartoszko JJ, Zeraatkar D, et al. Drug treatments for covid-19: living systematic review and network meta-analysis. *BMJ* 2020;370:m2980. doi:10.1136/bmj.m2980.
16 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.
17 Ye Z, Rochwerg B, Wang Y, et al. Treatment of patients with nonsevere and severe coronavirus disease 2019: an evidence-based guideline. *CMAJ* 2020;192:E536-45. doi:10.1503/cmaj.200648.
18 Hultcrantz M, Rind D, Akl EA, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol* 2017;87:4-13. doi:10.1016/j.jclinepi.2017.05.006.

Supplementary information: Appendix 1
Supplementary information: Appendix 2
Supplementary information: Appendix 3