J Bras Pneumol. 2023;49(4):e20230167
https://dx.doi.org/10.36416/1806-3756/e20230167

SPECIAL ARTICLE

Certainty of evidence, why?

João Pedro Lima[1], Xiajing Chu[1], Gordon H Guyatt[1], Wimonchat Tangamornsuksan[1,2]

1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton (ON) Canada.
2. Princess Srisavangavadhana College of Medicine, Chulabhorn Royal Academy, Bangkok, Thailand.

Submitted: 11 May 2023.
Accepted: 12 May 2023.

Study carried out in the Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton (ON) Canada.

ABSTRACT

Optimal clinical decision-making requires understanding of evidence regarding benefits, harms, and burdens of alternative management options. Rigorously conducted systematic reviews and meta-analyses offer accurate summaries of the evidence. However, such summaries may review only low-certainty evidence, in the process highlighting that no single decision is likely to be best for all patients. The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach offers a systematic and transparent method for rating certainty of evidence in systematic reviews. In this paper, we will address the importance of assessing the certainty associated with bodies of evidence; explain how the GRADE system rates the certainty of evidence from systematic reviews; and present the GRADE evidence to decision framework for moving from evidence to strong or weak recommendations in clinical practice guidelines.

Keywords: Systematic reviews as topic; Meta-analysis as topic; Evidence-Based Medicine; Decision making.

INTRODUCTION

When answering patient questions regarding treatment options, clinicians need to consider the relevant evidence regarding benefits, harms, and burdens. Systematic reviews and meta-analyses address structured clinical questions and, when done well, offer accurate summaries of the evidence. When the evidence is low certainty (also known as low quality), however, even rigorous evidence summaries will leave large uncertainty regarding benefits and harms. In this paper, we will address the importance of assessing the certainty of the evidence from interventional and diagnostic studies and explain the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach to rating the certainty of evidence from systematic reviews and the strength of recommendations in clinical practice guidelines.

Patients, clinicians, and policymakers will often be misled if they do not consider the certainty of evidence. Consider the use of systemic glucocorticoids, until recently widely used in the management of idiopathic pulmonary fibrosis. The evidence supporting the benefit of glucocorticoid use in these patients was never better than low certainty, whereas high-certainty evidence exists for the multiple harms of this intervention.[1] Optimal practice for clinicians offering glucocorticoid therapy to patients would include making clear the speculative nature of any benefits and the high risk of substantial harm. Many patients, aware of the uncertain benefits and the high-certainty evidence of harms, would decline the intervention. Failure to recognize the low-certainty evidence of benefit would result in overuse of the intervention.

A formal assessment of the certainty of evidence is an effective strategy to prevent the overuse of interventions with questionable benefits. The GRADE approach offers a systematic and transparent method for rating certainty of evidence in systematic reviews (Chart 1), and for developing and determining the strength of recommendations in clinical practice guidelines.[2] More than 110 organizations, including the World Health Organization, the UK National Institute for Health and Care Excellence, the Cochrane Collaboration, and leading American professional organizations including the American Thoracic Society and the American College of Chest Physicians have adopted GRADE. Moreover, the world's leading electronic textbook, UpToDate, includes over 10,000 GRADE recommendations. GRADE now represents the gold standard approach to systematic reviews and guideline development.[3]

Applying the GRADE system of rating certainty of evidence requires the availability of rigorously conducting systematic reviews to address clinical questions. GRADE also offers evidence to decision (EtD) frameworks for guideline panels as they move from evidence to recommendations.[4] After considering all issues highlighted in the EtD framework, guideline panels will issue, in favor or against a treatment or diagnostic test, a strong or weak recommendation.

Naïve clinicians may be prematurely inclined to change their practice based on the results of a single randomized trial, neglecting considerations of risk of bias, imprecision due to limited sample size, and applicability if patients enrolled do not represent a close match to the patients under their care. Moreover, naïve clinicians may be

Correspondence to:
Wimonchat Tangamornsuksan. Princess Srisavangavadhana College of Medicine, Chulabhorn Royal Academy, 10210, Bangkok, Thailand. Tel.: 66 838704088.
E-mail: tangamow@mcmaster.ca or wimonchat.tan@gmail.com; or João Pedro Lima. Department of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main Street West 2C Area, Hamilton, Ontario, L8S 4K1, Canada. Tel.: 1 365 888-0631. E-mail: limaj1@mcmaster.ca
Financial support: None.

Chart 1. Certainty of evidence: assessment criteria.

Study design	Confidence in estimates	Lower if	Higher if
Randomized trials	High	Risk of bias -1 Serious -2 Very serious	Large effect +1 Large +2 Very large
	Moderate	Inconsistency -1 Serious -2 Very serious	Dose response +1 Evidence of a gradient
Observational studies	Low	Indirectness -1 Serious -2 Very serious	All plausible confounding +1 Would reduce a demonstrated effect or
	Very low	Imprecision -1 Serious -2 Very serious Publication bias -1 Likely -2 Very likely	+1 Would suggest a spurious effect when results show no effect

ready to inappropriately change practice based on a systematic review and meta-analysis that yields only low-certainty evidence. The evidence may be low certainty if it comes exclusively from observational, non-randomized studies. Alternatively, the evidence may be low certainty, even if based on randomized trials, if those trials suffer from limitations in the study design and sample size; inconsistency in results; or limitations in applicability to the patients at hand. In the following sections of this review, we will expand on these limitations of randomized controlled trials (RCTs) and systematic reviews and meta-analyses in the clinical decision-making process, highlighting the importance of GRADE for rating the certainty of evidence and recommendations of treatment and diagnostic tests in clinical practice guidelines.

THE GRADE APPROACH IN SYSTEMATIC REVIEWS AND CLINICAL RECOMMENDATIONS

GRADE approach for rating certainty of evidence regarding interventions

The GRADE approach to the certainty of evidence begins with the acknowledgment that sound clinical decisions require rigorous systematic summaries of the highest quality available evidence regarding interventions under consideration. Once such a systematic review is available, the GRADE rating of the certainty of evidence begins with the study design: randomized trials begin as high-certainty evidence and observational studies as low-certainty evidence in GRADE's four-category system of certainty of evidence (high, moderate, low, and very low; Chart 1).[2] Following the study design, GRADE has identified five domains that warrant consideration when rating the certainty of evidence: risk of bias, inconsistency, indirectness, imprecision, and publication bias (Chart 1).[2]

Reviewers rate down the certainty of evidence by one level when they identify serious concerns and by two levels when they identify very serious concerns in any of these five domains. Reviewers can rate up the certainty of evidence from observational studies, primarily for large or very large magnitude of effect.[5] Reviewers assess the certainty of evidence not for individual studies but rather for entire bodies of evidence summarized in systematic reviews, and separately for each outcome. All patient-important outcomes receive a certainty rating.

We will now briefly describe considerations related to the five reasons for rating down the certainty of evidence. Concerning the risk of bias,[6] randomized trials may be limited by failure to conceal randomization; failure to blind patients, clinicians, data collectors, and adjudicators; and losing patients to follow-up. Randomized trials will also overestimate treatment effects if they are stopped early for large treatment effects, particularly if their sample size is small.[7]

Secondly, certainty decreases when there is unexplained inconsistency among results presented from different studies. Reviewers judge consistency through the similarity of point estimates and the extent of overlap of CIs. Statistical criteria may further inform judgments regarding inconsistency, including tests of heterogeneity (Can chance explain differences in results between studies?) and I^2, which quantifies inconsistency on a scale from 0 to 100.[8,9]

Thirdly, studies included in a systematic review should reflect the review question. When rating indirectness (the GRADE term related to the applicability of the evidence to the question at hand), reviewers consider whether patients, interventions, comparisons, and outcomes differ from those of interest.[10] Indirectness is even more important for guidelines than for systematic reviews.

Fourthly, GRADE considers the width of the CIs around the estimates of the absolute effects of treatment.[11] Rating down the certainty of evidence requires consideration of whether the CI crosses a threshold of interest. For instance, if the entire confidence is in the range of an important effect, one will not rate down for imprecision. If it crosses the threshold of importance, leaving uncertainty about whether an effect is trivial or important, reviewers will rate it down. Consider for example Figure 1: for intervention A, reviewers would rate down for imprecision, whereas, for intervention B, they would not.

Finally, trials that fail to show positive treatment effects may remain unpublished and thus result in overestimates of treatment effect, a phenomenon referred to as publication bias.[12] Review authors will suspect publication bias when a pharmaceutical company has sponsored all available studies, particularly if the sample size of the studies is small.

If a body of evidence from randomized trials suffers from several of these limitations, reviewers may rate down to moderate, low, or even very low certainty of evidence. Moreover, these limitations also apply to observational studies and may lead to rating down certainty from low to very low. On rare occasions, reviewers may rate up certainty for large or very large effects (e.g., insulin for diabetic ketoacidosis and dialysis for end-stage renal disease).

As with therapeutic interventions, systematic reviews should inform diagnostic clinical questions.[13] Most studies of diagnostic tests focus exclusively on diagnostic accuracy, and GRADE's five reasons for rating down apply to systematic reviews of such studies.[14] Ideally though, studies will focus on the impact of alternative diagnostic strategies on patient-important outcomes (e.g., mortality and quality of life) using randomized study designs.[15,16] For those studies, the certainty of evidence is assessed in the same way as the GRADE approach to clinical interventions.

How does GRADE inform moving from evidence to recommendations?

GRADE uses the EtD framework to help people use the evidence to inform clinical decisions. This framework includes considerations of the magnitude of benefits, harms, and burdens; the certainty of evidence regarding those benefits, harms, and burdens; patient values and preferences; and, sometimes, costs, feasibility, acceptability, and equity issues (Chart 2).[4] Clinical recommendations, after considering all these issues, should provide explicit statements on the best course of action.

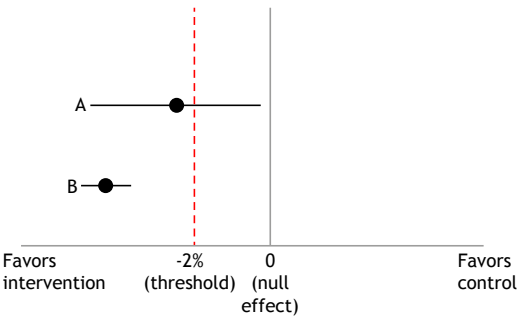

Figure 1. Rating imprecision: consideration of whether the confidence interval crosses a threshold of interest.

Guideline panels make strong recommendations when they conclude that all or almost all fully informed patients would choose the proposed intervention. In contrast, they make weak (also referred to as conditional) recommendations when they consider that patients presented with the treatment options would, as a result of different values and preferences, vary in their choices.[17]

Desirable and undesirable outcomes (estimated effects)

When benefits (desirable outcomes) are large, and harms and burdens (undesirable outcomes) are small in magnitude, guideline panels are more likely to issue a strong recommendation. In contrast, when the desirable and undesirable consequences are closely balanced, a weak recommendation is likely more appropriate.

Certainty of evidence

When evidence certainty is high or moderate, strong recommendations may be appropriate. When the evidence is low or very low certainty, high confidence that benefits outweigh harms and burdens (or the reverse) is very unlikely, and weak recommendations will almost always be appropriate.

Uncertainty or variability in values and preferences

Marking a recommendation involves determining the value one places on benefits versus harms and burdens. Although patients will have different views regarding these values, in making recommendations guideline panels must focus on typical or average patient values and preferences. Given this is the case, large variability in values and preferences in the relevant patient population will make a weak recommendation more likely, as will uncertainty regarding patient values and preferences. Although there is often limited evidence to inform patient preferences and values, clinical experience may leave a panel confident that values and preferences differ widely among patients.[14]

Resource use (costs), feasibility, acceptability, and equity

Treatment interventions or diagnostic tests may increase or decrease resource use when compared to an alternative. The impact of the cost may vary among settings and patients' socioeconomic situations. Additional, often secondary, considerations include resource use, feasibility, acceptability, and equity. Although these considerations are not always germane, they are sometimes important, particularly when guidelines take a public health or systems perspective rather than an individual patient perspective.

Chart 2. Domains that affect the strength of a recommendation.

- Desirable and undesirable outcomes (estimated effects)
- Certainty of evidence
- Uncertainty or variability in values and preferences
- Resource use (cost), feasibility, acceptability, and equity

How do clinicians interpret and apply GRADE recommendations to patient care?

Clinicians should be able to differentiate an untrustworthy recommendation from trustworthy recommendations; understand the meaning of the strength of the recommendation; and understand how to apply the recommendation to patient care.[18] A guide for health professionals to interpret and use recommendations in guidelines developed with the GRADE approach suggests specific criteria to interpret, critically assess, and apply GRADE recommendations (Chart 3).[17]

Understanding the meaning of the strength of the recommendation

Clinicians' interpretation of GRADE recommendations should include consideration of the strength of the recommendation and the certainty of the evidence. Guideline panels using the GRADE approach will issue either strong or weak/conditional recommendations. If a guideline panel is confident that desirable effects outweigh undesirable consequences, they will issue a strong recommendation, usually framed as "we recommend."[17] On the other hand, if the guideline panel is less confident about the balance between desirable and undesirable consequences in the proposed course, they issue a weak recommendation, usually framed as "we suggest."

Panels issue weak recommendations when they believe that the recommendation is unlikely to apply to all patients. In that case, clinicians should spend time to ensure that each patient receives the therapeutic option that reflects their values and preferences.[19]

Distinguishing between trustworthy and untrustworthy recommendations

Clinicians should not only understand the concepts of strength of the recommendation and certainty of the evidence but should also be able to choose trustworthy guidelines to inform their practice. Consideration of five domains may help in this choice (Chart 3).[17]

Were all of the relevant outcomes important to patients explicitly considered?

Balancing between desirable and undesirable in the proposed course will depend on what outcomes are considered. Clinicians should assess whether the guideline panel considered and included all relevant patient-important outcomes.

Was the recommendation based on the best current evidence?

The recommendation should be based on the best current evidence. Clinicians should assess the credibility of the guideline process based on whether a systematic review informed the recommendations. Ideally, the systematic review panels should be up to date.

Is the strength of the recommendation appropriate?

Guideline panels should consider all issues in the EtD framework in making their recommendations and seldom make strong recommendations when evidence is low certainty (Chart 2).

Is the recommendation clear and actionable?

The recommendation should provide the details of the recommended action, the situation to which the recommendations apply, to whom they apply, and the clinical action to which the intervention was compared.

Applying recommendations to patient care

Clinicians can apply strong recommendations to all or almost all patients without the necessity of a detailed discussion with the patient. For weak recommendations, clinicians should understand and be able to communicate the evidence to patients through shared decision-making.

FINAL CONSIDERATIONS

Neither individual RCTs nor systematic reviews of the best available evidence ensure high-certainty evidence; indeed, for RCTs and rigorous systematic reviews, the certainty of the evidence may be low. The GRADE approach offers a system for rating the certainty of evidence in systematic reviews and grading the strength of recommendations in clinical guidelines. In applying guidelines to clinical care, clinicians should understand the implications of strong and particularly weak recommendations that mandate considering

Chart 3. User guide to GRADE for health professionals, including interpretation, critical assessment, and use of GRADE recommendations in patient care.

Understanding the meaning of the strength of the recommendation
What does strength mean?
What does the certainty of the evidence mean?
Distinguishing between trustworthy and untrustworthy recommendations
Were all of the relevant outcomes important to patients explicitly considered?
Was the recommendation based on the best current evidence?
Is the strength of the recommendation appropriate?
Is the recommendation clear and actionable?
Does the recommendation provide the necessary additional information?
Applying recommendations to patient care
Strong recommendations
Weak recommendations

GRADE: Grading of Recommendations Assessment, Development, and Evaluation.

patient values and preferences in their decision-making process.

AUTHOR CONTRIBUTIONS

JPL, XC, and WT equally contributed to this work. JPL, XC, GHG, and WT: study conception and design. JPL, XC, and WT: drafting of the manuscript. JPL, XC, GHG, and WT: review and approval of the final version of the manuscript.

CONFLICTS OF INTEREST

None declared.

REFERENCES

1. Pitre T, Mah J, Helmeczi W, Khalid MF, Cui S, Zhang M, et al. Medical treatments for idiopathic pulmonary fibrosis: a systematic review and network meta-analysis. Thorax. 2022;77(12):1243-1250. https://doi.org/10.1136/thoraxjnl-2021-217976
2. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-394. https://doi.org/10.1016/j.jclinepi.2010.04.026
3. GRADE Working Group [homepage on the Internet]. GRADE. Available from: https://www.gradeworkinggroup.org/
4. Alonso-Coello P, Schünemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. BMJ. 2016;353:i2016. https://doi.org/10.1136/bmj.i2016
5. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64(12):1311-1316. https://doi.org/10.1016/j.jclinepi.2011.06.004
6. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence–study limitations (risk of bias). J Clin Epidemiol. 2011;64(4):407-415. https://doi.org/10.1016/j.jclinepi.2010.07.017
7. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. JAMA. 2005;294(17):2203-2209. https://doi.org/10.1001/jama.294.17.2203
8. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence–inconsistency. J Clin Epidemiol. 2011;64(12):1294-1302. https://doi.org/10.1016/j.jclinepi.2011.03.017
9. Guyatt G, Zhao Y, Mayer M, Briel M, Mustafa R, Izcovich A, et al. GRADE guidance 36: updates to GRADE's approach to addressing. J Clin Epidemiol. 2023;158:70-83. https://doi.org/10.1016/j.jclinepi.2023.03.003
10. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence–indirectness. J Clin Epidemiol. 2011;64(12):1303-1310. https://doi.org/10.1016/j.jclinepi.2011.04.014
11. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence–imprecision [published correction appears in J Clin Epidemiol. 2021 Sep;137:265]. J Clin Epidemiol. 2011;64(12):1283-1293. https://doi.org/10.1016/j.jclinepi.2011.01.012
12. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence–publication bias. J Clin Epidemiol. 2011;64(12):1277-1282. https://doi.org/10.1016/j.jclinepi.2011.01.011
13. Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy. 2009;64(8):1109-1116. https://doi.org/10.1111/j.1398-9995.2009.02083.x
14. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies [published correction appears in BMJ. 2008 May 24;336(7654). doi: 10.1136/bmj.a139. Schünemann, A Holger J [corrected to Schünemann, Holger J]]. BMJ. 2008;336(7653):1106-1110. https://doi.org/10.1136/bmj.39500.677199.AE
15. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ. 1986;134(6):587-594.
16. El Dib R, Tikkinen KAO, Akl EA, Gomaa HA, Mustafa RA, Agarwal A, et al. Systematic survey of randomized trials evaluating the impact of alternative diagnostic strategies on patient-important outcomes. J Clin Epidemiol. 2017;84:61-69. https://doi.org/10.1016/j.jclinepi.2016.12.009
17. Neumann I, Santesso N, Akl EA, Rind DM, Vandvik PO, Alonso-Coello P, et al. A guide for health professionals to interpret and use recommendations in guidelines developed with the GRADE approach. J Clin Epidemiol. 2016;72:45-55. https://doi.org/10.1016/j.jclinepi.2015.11.017
18. Brignardello-Petersen R, Carrasco-Labra A, Guyatt GH. How to Interpret and Use a Clinical Practice Guideline or Recommendation: Users' Guides to the Medical Literature [published correction appears in JAMA. 2022 Feb 22;327(8):784]. JAMA. 2021;326(15):1516-1523. https://doi.org/10.1001/jama.2021.15319
19. Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol. 2013;66(7):719-725. https://doi.org/10.1016/j.jclinepi.2012.03.013