available at www.sciencedirect.com
journal homepage: www.europeanurology.com/eufocus

Mini Review

Evaluating the Certainty of Evidence in Evidence-based Medicine

Daniel A. González-Padilla [a,*], *Philipp Dahm* [b,c]

[a] *Department of Urology, Clínica Universidad de Navarra, Madrid, Spain;* [b] *Minneapolis VA Healthcare System, Minneapolis, MN, USA;* [c] *Department of Urology, University of Minnesota School of Medicine, Minneapolis, MN, USA*

Article info

Article history:
Accepted October 11, 2023
Available online 21 October 2023
Associate Editor: Christian Gratzke

Keywords:
GRADE framework
Evidence
Certainty
Quality

Abstract

Certainty of evidence (formerly known as quality of evidence) is defined as the extent to which our confidence in an estimate of the effect is correct or our certainty that such estimate supports a particular recommendation for a clinical practice guideline. Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) is a structured and reproducible framework for assigning a level of certainty on a per-outcome basis for evidence derived from randomized and nonrandomized studies. The level of certainty starts as high or low and can be increased or decreased after considering several criteria (eg, risk of bias, inconsistency of results, publication bias, dose-response gradient, large magnitude of effect, among others). Here we describe in brief the GRADE process for summarizing and assigning a certainty rating for evidence.

Patient summary: The GRADE framework is a way to work out how much we can trust results from medical research studies. This helps doctors in making informed decisions with their patients.

1. Introduction

One of the paradigms of evidence-based medicine is that clinical and health decision-making should be based on the current best evidence [1]. To identify what the best evidence is, we need a framework that allows us to assess how much we can trust a given body of evidence (ie, the studies identified as being related to a research question). While several such frameworks exist, GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) is the most widely used approach that is both methodologically rigorous and transparent; an introduction to the GRADE approach has been covered in a previous review [2]. GRADE defines the certainty of evidence (CoE; formerly known as "quality of evidence") as reflecting "the extent of our confidence that the estimates of the effect are correct" in the context of a systematic review, and as "the extent of our confidence that the estimates of an effect are adequate to support a particular decision or recommendation" in the context of a clinical practice guideline. While recognizing that such confidence exists on a continuum, for practical reasons, GRADE proposes four categories (Table 1). In a systematic review, CoE is determined on a per-outcome basis; in a guideline, CoE should qualify each recommendation, since each recommendation impacts multiple outcomes. GRADE is increasingly being used for systematic reviews in urology [3]. Among the urology-relevant guidelines, GRADE is used to rate the CoE in the American

* Corresponding author. Department of Urology, Clínica Universidad de Navarra, C. del Marquesado de Sta. Marta, 1, 28027 Madrid, Spain. Tel. +34 6 56565183.
E-mail address: dgonzalezp@unav.es (D.A. González-Padilla).

https://doi.org/10.1016/j.euf.2023.10.014

Table 1 – Categories for the certainty of evidence (taken from the GRADE handbook [11])

Grade	Definition
High	We are very confident that the true effect lies close to that of the estimate of the effect.
Moderate	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect.
Very low	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of the effect

Urological Association guidelines with the modification that the "low" and "very low" categories have been collapsed. GRADE is also used in selected European Association of Urology [4,5] and Canadian Urological Association [6] guidelines. In the following, we describe the GRADE process for assessing the CoE, as depicted in Figure 1.

2. Determining the CoE: rating down

GRADE assumes that a body of evidence from randomized controlled trials (RCTs) and nonrandomized studies (NRS) start off as evidence of high and low certainty, respectively. To what extent this is true depends on to what extent certain criteria are met, which might prompt us to lower or raise our confidence (not all RCTs are of high quality, and not all observational studies are of low quality).

GRADE comprises five domains for rating down, which apply to both RCTs and NRS. Substantial concerns for each domain may prompt us to lower our confidence and rate down the CoE. First, the study limitations domain is best known, since it corresponds to the risk-of-bias assessment that has long been part of any critical appraisal process. Examples include concerns regarding patients selection (via lack of allocation concealment) or performance bias (due to lack of blinding) [7].

Second, the inconsistency domain refers to the extent to which the various studies included in a systematic review addressing a given PICO (Patient, Intervention, Comparison, Outcome) question yield similar (or different) results. If the results from the individual studies differ to a substantial degree that would not be expected on the basis of chance and cannot be explained (eg, via a subgroup analysis), GRADE states that this lowers confidence in a given effect-size estimate and should prompt rating down of the CoE.

Third, the imprecision domain refers to the width of the confidence interval (CI) and the number of events (measured as the optimal information size) [8]. If the CI for a

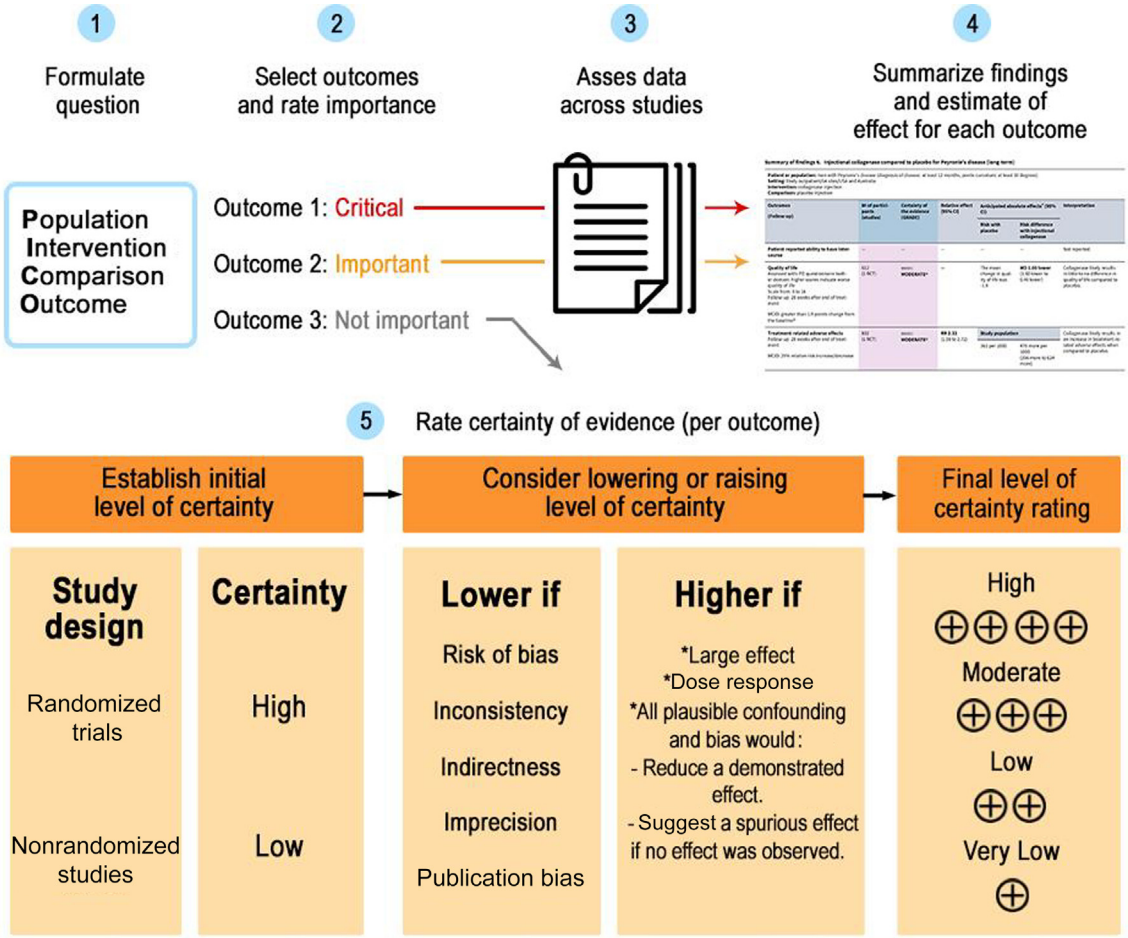

Fig. 1 – Schematic view of the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach for summarizing evidence and assigning a certainty of evidence rating (adapted from the GRADE handbook) [11].

given effect size crosses an assumed threshold of clinical importance, GRADE suggests rating down for imprecision (eg, a value of >3 is widely accepted as the threshold for a clinically meaningful change in International Prostate Symptom Score in men with lower urinary tract symptoms related to benign prostatic obstruction).

Fourth, the indirectness domain refers to the extent to which the evidence found corresponds to the PICO of interest. For example, authors of a systematic review may rate down the CoE for effect estimates of disease-specific survival in metastatic renal carcinoma if they only find information on progression-free survival as a surrogate outcome.

Lastly, the fifth domain is publication bias, which refers to the well-documented phenomenon that "negative studies" (which fail to demonstrate that a new procedure or drug works) are less likely to be published or are published in a delayed fashion in less well-known journals, therefore posing the risk that they are systematically omitted and thus biasing the analysis to showing a greater effect than may be true [9].

3. Determining the CoE: rating up

In the current GRADE framework, rating up only applies to NRS (that start as low CoE), with no additional serious limitations (eg, a large proportion of participants being lost to follow-up) and is therefore a relatively infrequent event. GRADE has defined three reasons that may prompt rating up under specific circumstances. First and most relevant is the domain of magnitude of effect, meaning that our confidence may increase when an NRS shows a large or very large effect size. The underlying rationale is that such effect sizes are unlikely to be entirely explained by bias alone, therefore increasing our confidence that at least part of the effect is "real". For example, we have major confidence in the ability of medical (or surgical) castration to relieve bone pain in patients with metastatic, treatment-naïve prostate cancer, although this is based solely on NRS evidence. GRADE therefore suggests considering rating up the CoE by one level when methodologically rigorous observational studies show at least a twofold change in risk, and rating up by two levels for at least a fivefold reduction change in risk.

Second is the dose-response domain, related to the finding that an increase in the magnitude of an intervention or exposure corresponds to an increase in the change for the outcome. For example, a systematic review of NRS may demonstrate that a decrease in renal function corresponds closely to the amount of functional renal parenchyma removed at the time of partial nephrectomy.

Finally, there is the rare scenario in which there is likely to be residual confounding that is unaccounted for but is likely to further strengthen the association observed rather than weaken it. For example, if the results from a well-conducted systematic review of observational studies that compared prostate cancer outcomes between for-profit and not-for-profit hospitals favored the not-for-profit hospitals without adjusting for cancer stage (likely to be more advanced) and overall resources (likely to be less), this potential residual confounding may strengthen our confidence in the association observed and prompt us to rate the CoE upwards [10].

4. Conclusions

The GRADE CoE concept provides users of the medical literature with an assessment of how much confidence they can place in a given result, and it should therefore be used to qualify every effect-size estimate and guideline recommendation.

Conflicts of interest: The authors have nothing to disclose.

References

[1] Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71–2. https://doi.org/10.1136/bmj.312.7023.71.
[2] Gonzalez-Padilla DA, Dahm P. Evidence-based urology: understanding GRADE methodology. Eur Urol Focus 2021;7:1230–3. https://doi.org/10.1016/j.euf.2021.09.014.
[3] Norling B, Jung JH, Hwang EC, et al. GRADE reporting in systematic reviews published in the urological literature (2009–2021). J Urol 2023;210:529–36. https://doi.org/10.1097/JU.0000000000003558.
[4] Tikkinen KAO, Cartwright R, Gould MK, et al. EAU guidelines on thromboprophylaxis in urological surgery. Arnhem, The Netherlands: European Association of Urology; 2017.
[5] Dahm P, Cleveland B, Lauwagie A, Gonzalez-Padilla DA. Adherence of the European Association of Urology guidelines to the Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology. Eur Urol. In press. https://doi.org/10.1016/j.eururo.2023.02.023.
[6] Richard PO, Violette PD, Bhindi B, et al. 2023 update – Canadian Urological Association guideline: management of cystic renal lesions. Can Urol Assoc J 2023;17:162–74. https://doi.org/10.5489/cuaj.8389.
[7] Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898. https://doi.org/10.1136/bmj.l4898.
[8] Zeng L, Brignardello-Petersen R, Hultcrantz M, et al. GRADE Guidance 34: update on rating imprecision using a minimally contextualized approach. J Clin Epidemiol 2022;150:216–24. https://doi.org/10.1016/j.jclinepi.2022.07.014.
[9] Tseng TY, Stoffs TL, Dahm P. Evidence-based urology in practice: publication bias. BJU Int 2010;106:318–20. https://doi.org/10.1111/j.1464-410X.2010.09380.x.
[10] Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011;64:1311–6. https://doi.org/10.1016/j.jclinepi.2011.06.004.
[11] Schünemann H, Brożek J, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. The GRADE Working Group; 2013.