Journal of Clinical Epidemiology 198 (2026) 112411

OTHER GRADE PAPER

Artificial Intelligence-enabled GRADE: how the GRADE Working Group will use automation to rate the certainty of evidence of intervention effects

Bernardo Sousa-Pinto[a,b], Manuel Marques-Cruz[a,b], Rafael José Vieira[a,b], Vítor Henrique Duarte[a,b], Camila Ávila-Oliver[c], Andreea Dobrescu[d], Paweł Jemioło[e,f], Arnav Agarwal[g], Malgorzata M. Bala[f], Jessica Beltran[h,i], Antonio Bognanni[j], Fabio Cruciani[k], Omar Dewidar[l,m], Sara Gil-Mata[a,b], Changhao Liang[n,o], Daniel Martinho-Dias[a,b], Claus Nowak[d], Honoria Ocagli[p], Paula Perestrelo[a,b,q], Ana Carolina Pereira Nunes Pinto[h], Joana Reis-Pardal[a,b], Pau Riera-Serra[r], Aline Rocha[s,t], Karina Fernández-Saénz[h], Itziar Etxeandia-Ikobaltzeta[u], Curtis S. Harrod[u], Miranda W. Langendam[v], Joerg Meerpohl[w,x], Francesco Nonino[y], Wojtek Wiercioch[o,z], Ignacio Neumann[aa], Ariel Izcovich[ab,ac], Silvia Minozzi[ad], Gerald Gartlehner[d,ae], Holger J. Schünemann[af,ag,*]

[a]*Faculty of Medicine, MEDCIDS - Department of Community Medicine, Information and Health Decision Sciences, University of Porto, Porto, Portugal*
[b]*CINTESIS@RISE - Health Research Network, Faculty of Medicine, MEDCIDS, University of Porto, Porto, Portugal*
[c]*Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Santiago, Chile*
[d]*Cochrane Austria, Department for Evidence-Based Medicine and Evaluation, University for Continuing Education Krems, Krems, Austria*
[e]*AGH University of Krakow, Kraków, Poland*
[f]*Jagiellonian University Medical College, Krakow, Poland*
[g]*Division of General Internal Medicine, Department of Medicine, University of Alberta, Edmonton, Alberta, Canada*
[h]*Iberoamerican Cochrane Centre - Biomedical Research Institute Sant Pau (IIB Sant Pau), Barcelona, Spain*
[i]*Institut de Recerca Sant Pau (IR Sant Pau), Barcelona, Spain*
[j]*Department of Biomedical Sciences, Clinical Epidemiology and Research Center (CERC), Humanitas University, Pieve Emanuele, Milan, Italy*
[k]*Department of Epidemiology Lazio Regional Health Service - ASL Roma 1, Rome GRADE Centre, Rome, Italy*
[l]*Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada*
[m]*Bruyère Health Research Institute, University of Ottawa, Ottawa, Ontario, Canada*
[n]*Centre for Evidence-Based Chinese Medicine, Beijing University of Chinese Medicine, Beijing, China*
[o]*Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada*
[p]*Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences, and Public Health, University of Padova, Padova, Italy*
[q]*Local Health Unit of Trás-os-Montes e Alto Douro, Oncology Department, Vila Real, Portugal*
[r]*Health Research Institute of the Balearic Islands, Palma, Balearic Islands, Spain*
[s]*Universidade Federal de São Paulo (UNIFESP/EPM), Disciplina de Medicina de Urgência e Medicina Baseada em Evidências, São Paulo, São Paulo, Brazil*
[t]*Cochrane Brazil, Núcleo de Avaliação Tecnológica em Saúde, São Paulo, São Paulo, Brazil*
[u]*American College of Physicians, Philadelphia, PA, USA*
[v]*Amsterdam UMC, University of Amsterdam, Epidemiology and Data Science, Amsterdam Public Health Research Institute, Methodology Program, Amsterdam, The Netherlands*
[w]*Institute for Evidence in Medicine, Medical Center & Faculty of Medicine, University of Freiburg, Freiburg, Germany*
[x]*Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany*
[y]*IRCCS Istituto Delle Scienze Neurologiche di Bologna, Bologna, Italy*
[z]*Michael G. DeGroote Cochrane Canada & McMaster GRADE Centres, McMaster University, Hamilton, Ontario, Canada*

Funding: None.

* Corresponding author. Clinical Epidemiology and Research Center (CERC), Humanitas University and IRCSS Research Hospital, Via Rita Levi Montalcini 4, Pieve Emanuele, Milan 20090, Italy.
E-mail address: schuneh@mcmaster.ca (H.J. Schünemann).

https://doi.org/10.1016/j.jclinepi.2026.112411
0895-4356/

[aa]School of Medicine, Universidad San Sebastián, Santiago, Chile
[ab]Facultad de Medicina, Universidad del Salvador, Buenos Aires, Argentina
[ac]Servicio de Clínica Médica, Hospital Alemán, Buenos Aires, Argentina
[ad]Laboratory of Methodology of Systematic Reviews and Guidelines production, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy
[ae]Center for Health Economics, Methods, and Evidence Synthesis, RTI International, Research Triangle Park, NC, USA
[af]Charit, Allergology and Immunology, Universitätsmedizin-Berlin, Berlin, Germany
[ag]IRCCS Humanitas Research Hospital, Pieve Emanuele, Milan, Italy

Accepted 2 July 2026; Published online 7 July 2026

Abstract

Objectives: The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group is developing GRADErater (GRADE Rating Automation Through Enhanced Reasoning), an official, automated tool for evaluation of the certainty of evidence (CoE). In this article, we describe the principles and methods underlying its development and state how the GRADE Working Group (GWG) will use automation to rate the CoE of intervention effects.

Methods: We followed the GRADE methods for registering a project group on GRADE and artificial intelligence (AI). The project group established that the automated tool should be developed according to the following principles: (i) compliance with the current GRADE guidance, (ii) transparency, (iii) human oversight, (iv) implementation of decision rules, (v) ease of use, (vi) understandability, and (vii) continuous improvement. We developed a set of decision rules to appraise each domain of the CoE in pairwise and network meta-analysis. These rules were developed based on official GRADE sources and validated by GRADE experts. Based on these rules, we created a first version (https://gradeai.med.up.pt/) for GRADErater. We are now evaluating this version in terms of (i) its underlying rules and (ii) its user interface. These assessments will allow for a refinement of the rules and of the first version. The modified version will be appraised and presented to the GWG with input from internal and external interest-holders before we seek formal approval. Once launched, the automated tool will be continuously evaluated and refined by incorporating feedback from end users. We will add new features (including generative AI-based functionalities), integrate this tool with GRADEpro, and develop versions in languages other than English.

Conclusion: This project will follow a transparent methodology to create GRADErater, an official tool endorsed by the GWG that will support humans in applying the most current GRADE methods to rate the CoE.

Keywords: Artificial intelligence; Automation; Certainty of evidence; GRADE; Evidence synthesis; Network meta-analysis

1. Introduction

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) Working Group has developed a systematic and transparent approach for evaluating the certainty of evidence (CoE) (book.gradepro.org) that is used in evidence syntheses, health technology assessments, and health guidelines. The CoE is one of the criteria in the Evidence-to-Decision framework, an approach that allows for structured and transparent health decisions based on the best available evidence [1,2].

Using GRADE to appraise the CoE involves judging whether there are no serious, serious, very serious, or extremely serious concerns in defined domains. This can be a time-consuming process, in particular when appraising the CoE for network meta-analysis (NMA), as such implies evaluating multiple comparisons and in a larger set of domains [3,4]. Minozzi et al have reported median times of 160−481 minutes to evaluate the CoE of an NMA with GRADE [5]. Noteworthily, assessing the CoE of an NMA is a stepwise approach in which mistakes in one step can be carried forward. In this context, and even though applying GRADE involves a degree of context knowledge and subjectivity, automation can contribute to increased consistency and efficiency. Recent updates in the GRADE methodology have facilitated automation: the adoption of an approach based on multiple decision thresholds (reference estimates that allow to classify the magnitude of the effect size of an intervention as trivial or none, small, moderate, or large) [6,7] has enabled the establishment of a correspondence between the number of thresholds and the number of levels to rate down for each domain. In addition, the most current approaches to appraise the CoE have been systematized in the updated version of the GRADE book, GRADE's official resource for its application [8].

Therefore, the GRADE Working Group (GWG) through its GRADE for artificial intelligence (AI) project group has launched an initiative aiming at the creation of GRADErater (GRADE Rating Automation Through Enhanced Reasoning), GRADE's automated tool to support the GRADE assessment of the CoE. In this paper, we discuss the principles and methods we are applying in the development of such a tool, informing how GRADE will address automation and AI in the rating of the CoE obtained from meta-analysis of randomized controlled trials or

What is new?

Key findings

- The GRADE Working Group has launched an initiative aiming at creating GRADErater, an official automated GRADE tool to support the assessment of the certainty of evidence (CoE).
- The development of GRADErater will build on several principles, namely compliance with the most current GRADE guidance, transparency, human oversight, implementation of decision rules, ease of use, understandability, and continuous improvement.

What this adds to what was known?

- As part of an effort to make CoE assessment more efficient and transparent, the GRADE Working Group is developing an automated tool. A first version of this tool has been built and is undergoing testing in terms of underlying rules and interface.

What is the implication and what should change now?

- Once launched, GRADErater has the potential to make the CoE assessment more efficient and transparent.

nonrandomized studies. A subsequent publication will present the underlying decision-based rules that will have undergone validation and been approved by the GWG with the finalized, living tool.

2. Principles for automation

The automated tool was devised in such a way that, within few minutes, generates automated suggestions of judgments for the different GRADE domains by performing background calculations and applying decision rules. Therefore, GRADErater may result in reduced time to assess the CoE, overcoming one of the challenges of applying GRADE. However, prior to the development of GRADErater, we established a set of principles through discussion that this automated tool will need to comply. These principles are not focused on how automation will result in higher efficiency but have rather been set to ensure a responsible and ethical use of automation (although they will make rating more efficient), and are in line with the Guidelines International Network (GIN) principles on the use of automation and AI for guideline development [9]:

- **Compliance with the most current GRADE guidance**: The final automated tool will have the approval of the GRADE Guidance Group (GGG) and will be part of the GRADE's Working Group set of endorsed official tools and products. Therefore, one of the defined principles was that automation should comply with the most current GRADE methodology as described in the GRADE book [8] and the up-to-date GRADE guidance papers as defined by the GRADE living map [10].
- **Transparency:** Automation should be implemented in a way that ensures transparency and accountability by providing end users (i) all automation rules that have been defined in general and (ii) the justification for each individual automated judgment each time automation is used. That is, end users should not only have access to the suggested automated judgment for each CoE domain but also to the underlying justification.
- **Human oversight**: Automation is developed not to replace but rather to support humans in the evaluation of the CoE. That is, automated judgments should not be understood as definite assessments but rather as initial suggestions that need to be checked by qualified humans. This means that automation should include mechanisms that allow humans to verify, understand, and (if needed) modify the automated judgments.
- **Implementation of decision rules**: To comply with the ''transparency'' and ''human oversight'' principles, we opted for implementing automation based on deterministic decision rules instead of generative AI. For the specific tasks of suggesting ratings of the CoE, generative AI would have several disadvantages, including the possibility of hallucinations, limitations in explainability (ie, in providing justifications for the judgments), less predictability and consistency (ie, different judgments could end up being suggested for similar situations), and higher economic costs and environmental impact. Algorithms based on a set of predefined automatable decision rules are able to overcome all these challenges, ensuring predictable and consistent judgments that can be appraised in a more transparent way. That is, these decision rules may promote standardization of the judgments. Nevertheless, future iterations of the automated tool may include some components with generative AI, including for supporting data preparation or in the context of integration with the AI-based chatbot of GRADE book.
- **Ease of use:** Automation should result in a final product that guarantees increased efficiency and is easily usable by those conducting GRADE assessments (eg, methodologists and systematic review authors). This implies the development of an intuitive user interface that allows for the maximum number of

outputs based on the minimum required number of inputs.

- **Understandability:** GRADE users not only include GRADE methodologists but also (i) health professionals addressing specific health questions in practice and policy, such as clinicians, public health officers, and policymakers; and (ii) decision-makers, such as participants in guideline development groups, individuals making coverage decisions, or those conducting evidence syntheses, advising others or appraising GRADE outputs. Thus, beyond transparency and ease of use, simplified explanations about how the tool automates the GRADE assessment and on the meaning of results are essential for those users.
- **Continuous improvement:** Efforts should be made to ensure that the developed automation tool (i) is in line with future evolutions in GRADE methodology, (ii) integrates a growing number of relevant features, and (iii) can be applied beyond studies of intervention. In addition, channels should be established to receive feedback from end users on possible errors or bugs to be corrected or suggestions on how to improve the user interface.

3. Methods for tool development

For the development of an automated tool compliant with the aforementioned principles, four phases were established, namely: (i) development of the rules and expert evaluation; (ii) development and testing of the first version; (iii) launch of the automated tool; and (iv) continuous improvement. These four phases are depicted in Figure.

3.1. Phase 1: development of the rules and expert evaluation

Two researchers (BSP and RJV) reviewed official GRADE papers on the evaluation of the CoE for pairwise and NMA [10] and the GRADE book chapters [8]. Based on this review, they translated existing GRADE guidance into a set of decision rules to operationalize the appraisal of each CoE domain. That is, automation relies on decision rules that reflect GRADE guidance; each suggested judgment arises from explicit predefined rules and can be reviewed or overridden by users. The following principles were adopted for the development of such rules:

- The GRADE book is the official source for using the approach to appraise the CoE. It is based on up-to-date, peer-reviewed GRADE guidance and other GRADE-approved material. The decision rules for evaluating each domain of the CoE in the context of pairwise meta-analysis or the direct component of NMA were grounded on the content of the GRADE book [8].
- To be useful for decision-making, the basis of automated rating will be grounded on six decision thresholds defining the certainty ranges (large/moderate harms, moderate/small harms, small/trivial or no harms, trivial or no/small benefits, small/moderate benefits, and moderate/large benefits). For continuous outcomes, thresholds need to be directly introduced by the user in the tool (eg, for standardized mean differences, the traditionally used values of 0.2, 0.5, and 0.8 can be used [11]). For dichotomous outcomes, these thresholds for absolute effects can be empirically determined based on the utilities for the corresponding health state or directly introduced by the user in the tool [7]. Using more than two thresholds and following the GRADE EtD comes with the advantage of being able to define rules for rating down by more than one level which is required for automation [6]. We used GRADE's rules for rating down the CoE based on decision thresholds. For example, if the confidence interval of the fixed effects meta-analytical measure crosses two decision thresholds, the algorithm suggests rating down by two levels (ie, judging as ''very serious'') the CoE for imprecision.

The draft decision rules were presented and sent to a panel of senior and experienced GWG experts (IE-I, CSH, MWL, JM, FN, WW, IN, AI, SM, GG, and HJS) who provided feedback through online meetings and email discussion. Experts were selected on the basis of having wide experience of applying GRADE particularly in the use of GRADE in NMA. The rules were then modified to be consistent with GRADE guidance and sent again to the experts until consensus was reached. We required two rounds of feedback to reach consensus.

3.2. Phase 2: development and validation of the first version

Based on the rules agreed by the GRADE experts, four researchers (BSP, MMC, RJV, and VHD) developed a first version of GRADErater that can automatically generate suggestions for the CoE assessment in the context of pairwise and network meta-analyses. The user interface of the first version allows users to upload their dataset (identification, quantitative data, and risk of bias and indirectness assessment for each study), provide additional input (eg, setting decision thresholds or indicating the baseline risk), and view the resulting judgments. Behind the interface, the tool runs a set of programmed scripts that process the inputs and apply the predefined decision rules. The interface then displays the suggested judgment together with a brief text justification drafted by the tool. Users can

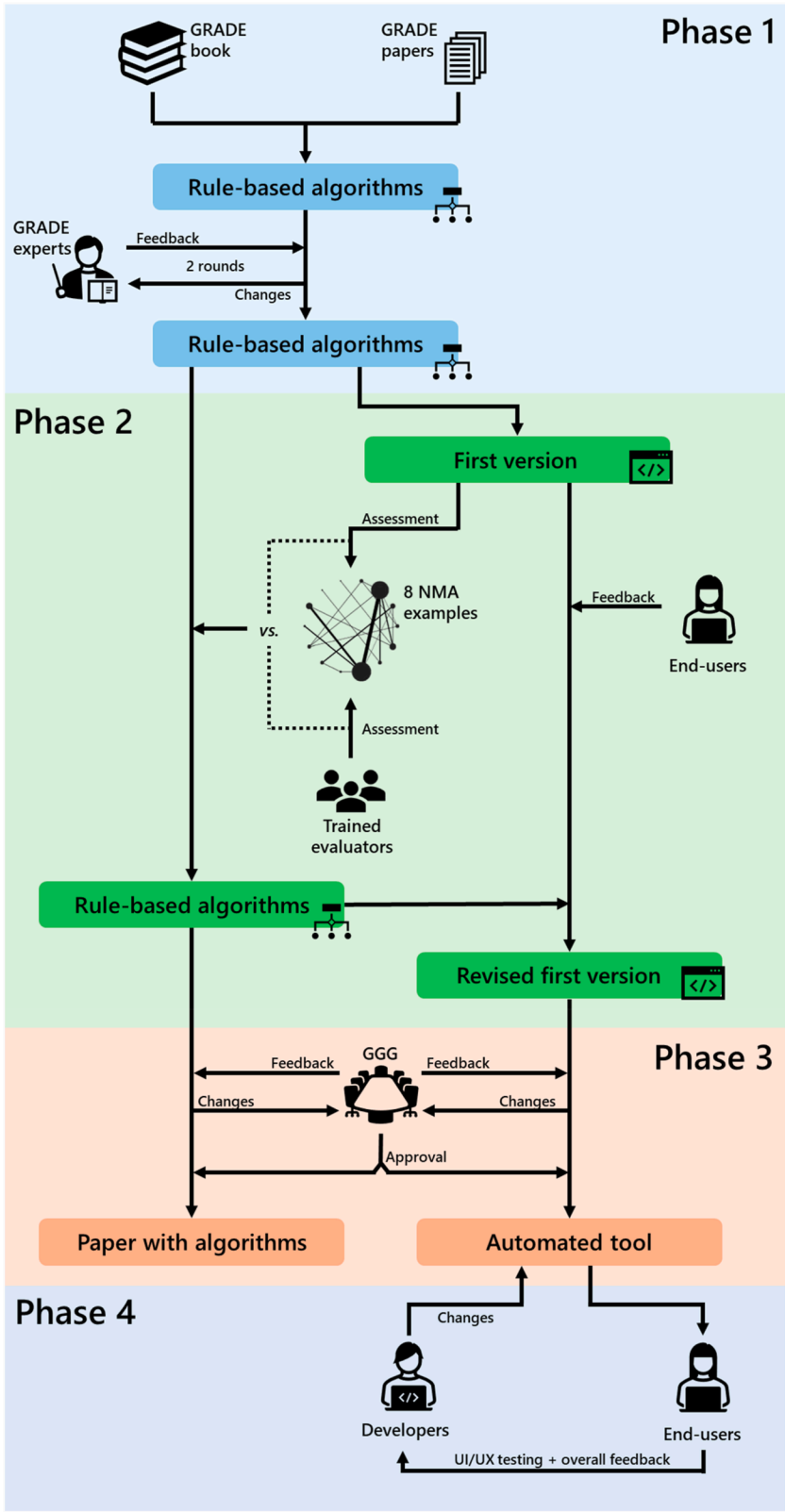

Figure. Graphical representation of the methods for the development of an automated tool supporting the GRADE evaluation of the CoE. GRADE, Grading of Recommendations Assessment, Development, and Evaluation; GGG, GRADE Guidance Group; NMA, network meta-analysis; UI/UX, user interface/user experience; CoE, certainty of evidence.

manually modify the suggested judgments if they do not agree with them. Additional outputs include the forest plots, network plots, net league tables, and intervention ranking tables. This first version is undergoing (i) testing to compare automated and human-only judgments based on examples and (ii) user interface/user experience testing.

A comparison between human-only and automated judgments is underway with the aims of identifying (i) eventual domains for which there is a systematic discrepancy between human-only and automated evaluations and (ii) eventual implementation errors. We opted to perform this comparison using examples of network meta-analyses. We chose this approach because appraising NMA includes evaluating CoE for direct pairwise components while considering a higher number of domains than in pairwise meta-analysis. In detail, we collected eight example network meta-analyses (ie, eight outcomes with multiple comparisons each) to test the developed first version: four were from a systematic review of intranasal treatments for allergic rhinitis [12] and the other four were from a systematic review on analgesics for migraine [13]. Moreover, five of the examples concerned dichotomous outcomes and three examples concerned continuous outcomes. The eight examples comprised 401 comparisons, of which there was direct evidence for 149. Each example will be independently evaluated without automation by two participants of the GRADE AI project group who had received training in appraisal of the CoE in the context of NMA (such training was ensured by attendance of three webinars given by HJS and BSP, which were then made available alongside standardized indications). Disagreements between independent human raters will be solved by consensus between them through online meetings, with registration of the reasons for disagreement; at least two online meetings are expected to occur: (i) one to discuss judgments on direct CoE domains, incoherence, and imprecision of indirect and NMA evidence and (ii) another to discuss the remaining domains (as their judgments are dependent on consensualized judgments regarding the direct CoE domains [3,4]). On the other hand, each example will be automatically assessed using the developed first version of GRADErater. One researcher will oversee the automated assessment of the example network meta-analyses. Since no modifications will be made to the automated suggested judgments of GRADErater, only one researcher will be needed to oversee those automated assessments. Researchers who oversee the automatic assessment are different from those involved in the human-only evaluations, and human-only assessments will be done blinded to the automated results. We will compute kappa coefficients and Gwet's AC1 to evaluate agreement between the consensualized human-only and automated judgments on each domain and in the overall CoE. Domains that will be particularly compared are those of direct CoE, incoherence, and imprecision of indirect and NMA evidence (the remaining domains are directly or indirectly dependent on the judgments for direct CoE domains). Following that comparison, discussion meetings will be held with the researchers involved in assessment of meta-analyses to explore potential reasons for the difference between the human-only and the automated judgments. The decision rules will then be reviewed, modified, and presented to the panel of GRADE experts until consensus is reached. Consensus is defined as all members of the expert group are in agreement following in-depth discussion in one or more online sessions, as necessary.

Prospective evaluations of GRADErater will also take place to register the time required for users to complete the tool, evaluate users' satisfaction with it in a ''real setting,'' and identify whether there are technical issues worth correcting.

In parallel, we will also be testing the user interface of the first version and user experience. In addition to formal testing involving participants recruited in GWG meetings and who are part of GRADE project groups, we will encourage potential users to provide their feedback. The first version of GRADErater can be found at https://gradeai.med.up.pt/. Importantly, the tool will only remain a first version for applying automation officially in GRADE until publication of a guidance article describing the implemented decision rules and integration with other official GRADE tools. Users can provide their feedback, which will then be discussed in the GRADE AI project group. Those who provide suggestions that are deemed worth implementation, following the International Committee of Medical Journal Editors criteria, will be invited to group authorship in the article describing the first completed version of the tool and the underlying rules.

The first version was created as Shiny web apps using R version 4.5.0 (R Foundation for Statistical Computing, Vienna, Austria), with pairwise and NMA being respectively implemented using meta and netmeta packages.

3.3. Phase 3: launch of the automated tool

The modified rules and the first version, reflecting the results of the comparisons with human-only assessments and the feedback provided in the prospective assessment to the user interface and user experience, will be presented to the GRADE AI project group and the GWG through email communications, regular presentations of the project group, and meetings of the GWG. We will iteratively incorporate the feedback of these groups until consensus is reached, and the automated tool is officially approved by the GGG. Once that happens, we will (i) launch GRADErater and (ii) draft a guidance article describing the tool and the underlying decision rules that will be voted for approval in a GWG meeting [14]. This article has been reviewed by

and received input from the GRADE Guidance Group in preparation of this future official GRADE guidance article.

3.4. Phase 4: continuous improvement

There will be efforts to ensure continuous improvement of the automated tool after its launch. Users will be able to send feedback directly to the developers, and project updates will be regularly sent to those interested through the GRADE newsletter and other dissemination tools. This will ensure an additional phase of workflow validation, with a larger pool of individuals with different backgrounds using, testing, and providing feedback on the tool.

In addition, we will integrate the tool with GRADEpro and continuously work on exploring the integration of new features, including AI-based modules to support the evaluation of risk of bias or indirectness and intransitivity, modules to support the assessment of potential dissemination bias, functions that support the computation of the net effect, an expansion to other contexts (eg, test accuracy), and full integration with the GRADE book. Therefore, the tool is expected to evolve, as it will include deterministic decision rules and generative AI components. The integration of the latter will take into account principles such as those set by GIN [9] or the Responsible use of AI in Evidence Synthesis [15]. Finally, we expect to release the tool in languages other than English to support overcoming language barriers.

4. Conclusion

Official GRADE guidance and the new version of the GRADE book (that include the adoption of multiple decision thresholds and the development of empirical methods to compute them) have laid conditions for the creation of a tool that can support the automated assessment of the CoE. By following a structured and transparent methodology, this project will create GRADErater, an official GRADE tool that will not replace human judgments but rather support humans in applying the most current methods to appraise the CoE. Although users may still need to manually evaluate the risk of bias and indirectness of each study initially, GRADErater will result in gains in efficiency, as it will automatically compute meta-analytical estimates, convert them into absolute measures (if applicable), apply decision thresholds, and suggest CoE judgments. GRADE will also automize risk of bias and indirectness judgments. In addition, GRADErater will enhance consistency and even transparency, with the rationale for each judgment being made explicit. This ensures an ethical and responsible use of automation that can support users incorporating the most recent GRADE developments and help them focusing on the more context-dependent aspects of the CoE assessment. The tool will be integrated with GRADE's software platform, GRADEpro. The first set of decision rules once approved by the GGG will be published in a GRADE guidance paper dedicated to the use of automation to rate the CoE.

Declaration of generative AI and AI-assisted technologies in the writing process

AI-assisted technologies were not used for data analysis, manuscript writing, figure production, or any other task.

Institutional review board approval statement

This study is exempt from IRB approval, considering its design.

CRediT authorship contribution statement

Bernardo Sousa-Pinto: Writing – original draft, Project administration, Methodology, Conceptualization. **Manuel Marques-Cruz:** Writing – original draft, Methodology, Conceptualization. **Rafael José Vieira:** Writing – original draft, Methodology, Conceptualization. **Vìtor Henrique Duarte:** Writing – original draft, Methodology, Conceptualization. **Camila Ávila-Oliver:** Writing – review & editing, Methodology. **Andreea Dobrescu:** Writing – review & editing, Methodology. **Paweł Jemioło:** Writing – review & editing, Methodology. **Arnav Agarwal:** Writing – review & editing, Methodology. **Malgorzata M. Bala:** Writing – review & editing, Methodology. **Jessica Beltran:** Writing – review & editing, Methodology. **Antonio Bognanni:** Writing – review & editing, Methodology. **Fabio Cruciani:** Writing – review & editing, Methodology. **Omar Dewidar:** Writing – review & editing, Methodology. **Sara Gil-Mata:** Writing – review & editing, Methodology. **Changhao Liang:** Writing – review & editing, Methodology. **Daniel Martinho-Dias:** Writing – review & editing, Methodology. **Claus Nowak:** Writing – review & editing, Methodology. **Honoria Ocagli:** Writing – review & editing, Methodology. **Paula Perestrelo:** Writing – review & editing, Methodology. **Ana Carolina Pereira Nunes Pinto:** Writing – review & editing, Methodology. **Joana Reis-Pardal:** Writing – review & editing, Methodology. **Pau Riera-Serra:** Writing – review & editing, Methodology. **Aline Rocha:** Writing – review & editing, Methodology. **Karina Fernández-Saénz:** Writing – review & editing, Methodology. **Itziar Etxeandia-Ikobaltzeta:** Writing – review & editing, Methodology. **Curtis S. Harrod:** Writing – review & editing, Methodology. **Miranda W. Langendam:** Writing – review & editing, Methodology. **Joerg Meerpohl:** Writing – review & editing, Methodology. **Francesco Nonino:** Writing – review & editing, Methodology. **Wojtek Wiercioch:** Writing – review & editing, Methodology. **Ignacio Neumann:** Writing – review & editing, Methodology. **Ariel Izcovich:**

Writing − review & editing, Methodology. **Silvia Minozzi:** Writing − review & editing, Methodology. **Gerald Gartlehner:** Writing − review & editing, Methodology. **Holger J. Schünemann:** Writing − review & editing, Project administration, Methodology, Conceptualization.

Declaration of competing interest

HJS is chair of the GRADE Working Group. The authors declare no direct financial conflicts of interest related to the content of this manuscript.

Acknowledgments

None.

Data availability

the grade rater is available online.

References

[1] Alonso-Coello P, Schunemann 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.

[2] Alonso-Coello P, Oxman AD, 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. 2: clinical practice guidelines. BMJ 2016;353:i2089. https://doi.org/10.1136/bmj.i2089.

[3] Brignardello-Petersen R, Florez ID, Izcovich A, Santesso N, Hazlewood G, Alhazanni W, et al. GRADE approach to drawing conclusions from a network meta-analysis using a minimally contextualised framework. BMJ 2020;371:m3900. https://doi.org/10.1136/bmj.m3900.

[4] Brignardello-Petersen R, Izcovich A, Rochwerg B, Florez ID, Hazlewood G, Alhazanni W, et al. GRADE approach to drawing conclusions from a network meta-analysis using a partially contextualised framework. BMJ 2020;371:m3907. https://doi.org/10.1136/bmj.m3907.

[5] Minozzi S, Cinquini M, Arienti C, Battain PC, Brigadoi G, Del Vicario M, et al. Grading of Recommendations Assessment, Development, and Evaluation and Confidence in Network Meta-Analysis showed moderate to substantial concordance in the evaluation of certainty of the evidence. J Clin Epidemiol 2025;184:111811. https://doi.org/10.1016/j.jclinepi.2025.111811.

[6] Schunemann HJ, Neumann I, Hultcrantz M, Brignardello-Petersen R, Zeng L, Murad MH, et al. GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions. J Clin Epidemiol 2022;150:225−42. https://doi.org/10.1016/j.jclinepi.2022.07.015.

[7] Wiercioch W, Morgano GP, Piggott T, Nieuwlaat R, Neumann I, Sousa-Pinto B, et al. GRADE guidance: using thresholds for judgments on health benefits and harms in decision making (GRADE guidance 42). Ann Intern Med 2025;178(11):1644−52. https://doi.org/10.7326/ANNALS-24-02013.

[8] GRADE Book. The GRADE Working Group. 2024. Available at: https://book.gradepro.org/ Accessed July 22,2026.

[9] Sousa-Pinto B, Marques-Cruz M, Neumann I, Chi Y, Nowak AJ, Reinap M, et al. Guidelines international network: principles for use of artificial intelligence in the health guideline enterprise. Ann Intern Med 2025;178(3):408−15. https://doi.org/10.7326/ANNALS-24-02338.

[10] Bognanni A, Sousa-Pinto B, McGowan J, Davoli M, Xia J, Skoetz N, et al. The GRADE Living Map (GLM): the interactive and living index of GRADE content. A statement by the GRADE Guidance Group. J Clin Epidemiol 2026:112419. https://doi.org/10.1016/j.jclinepi.2026.112419.

[11] Cohen J. Statistical power analysis for the behavioral sciences. In: . New York, NY: Routledge; 2013.

[12] Sousa-Pinto B, Vieira RJ, Bognanni A, Gil-Mata S, Ferreira-da-Silva R, Ferreira A, et al. Efficacy and safety of intranasal medications for allergic rhinitis: network meta-analysis. Allergy 2025; 80(1):94−105. https://doi.org/10.1111/all.16384.

[13] Gartlehner G, Dobrescu A, Wagner G, Chapman A, Persad E, Nowak C, et al. Pharmacologic treatment of acute attacks of episodic migraine: a systematic review and network meta-analysis for the American College of physicians. Ann Intern Med 2025;178(4): 507−24. https://doi.org/10.7326/ANNALS-24-02034.

[14] Schunemann HJ, Brennan S, Akl EA, Hultcrantz M, Alonso-Coello P, Xia J, et al. The development methods of official GRADE articles and requirements for claiming the use of GRADE - a statement by the GRADE Guidance Group. J Clin Epidemiol 2023;159:79−84. https://doi.org/10.1016/j.jclinepi.2023.05.010.

[15] Flemyng E, Noel-Storr A, Macura B, Gartlehner G, Thomas J, Meerpohl JJ, et al. Position statement on Artificial Intelligence (AI) use in evidence synthesis across cochrane, the Campbell Collaboration, JBI, and the Collaboration for environmental evidence 2025. Campbell Syst Rev 2025;21(4):e70074. https://doi.org/10.1002/cl2.70074.