Journal of PHYSIOTHERAPY
journal homepage: www.elsevier.com/locate/jphys

Appraisal

Clinimetrics: Grading of Recommendations, Assessment, Development and Evaluation (GRADE)

Summary

Description: The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) is a framework for assessing the quality (or certainty) of evidence and grading the strength of recommendations in healthcare.[1] The GRADE system has been endorsed by many organisations and it is becoming an international standard for use in judging the evidence in systematic reviews and clinical guidelines. This Clinimetrics paper only considers GRADE when used to judge evidence on treatment effects,[2] although the GRADE system can also rate the quality of evidence from diagnostic,[3] prognostic[4] and qualitative research.[5]

The quality of evidence is applied to each outcome and is rated at one of four levels: high, moderate, low and very low.[1] These levels imply a gradient of confidence in estimates of summary measures of treatment effect. Randomised controlled trials begin as high quality and observational studies begin as low quality. The confidence in the evidence can be decreased for five reasons: study limitations (risk of bias, such as lack of allocation concealment or blinding); inconsistency of results (heterogeneity or variability of point estimates and overlap of confidence intervals); indirectness of evidence (differences in populations, interventions, comparators or outcomes); imprecision of estimates (wide confidence intervals crossing a decision threshold); and publication bias (missing evidence, typically from studies that show no effect). Three situations could upgrade the quality of evidence: when the magnitude of the treatment effect is very large, if all plausible biases (confounding) would reduce a demonstrated effect and if there is evidence of a dose-response relationship.

GRADE provides two levels of recommendations – strong or weak – in favour of or against an intervention.[6] The strength of recommendation considers three factors: the balance between benefits and harms, variability in patients' preferences and values, and whether the intervention represents a wise use of resources. Online learning modules are available for the training of authors of systematic reviews and guideline developers, with most modules lasting no longer than 20 minutes.[7] The GRADE Working Group website also provides a list of publications and an online handbook for using GRADE and GRADEpro software.[8]

Clinimetric properties: In the first iteration of GRADE there was 'fair' inter-rater reliability for rating quality of evidence (kappa = 0.40).[2] After more guidance on how to use GRADE there was 'good' inter-rater reliability among inexperienced raters who received training on the GRADE methodology (intraclass correlation coefficient [ICC] = 0.66) and among members of the GRADE Working Group (ICC = 0.72).[9] The inter-rater reliability for quality of evidence has been shown to increase with training and with ratings by groups of three or four raters, but not when GRADE was assessed through a consensus rating.[9]

The inter-rater reliability for the individual GRADE domains was 'poor' to 'moderate' for risk of bias (kappa = 0.06 to 0.41), 'fair' to 'excellent' for inconsistency (kappa = 0.37 to 0.84), and 'poor' to 'moderate' for imprecision (kappa = 0.18 to 0.21), while for indirectness agreement varied between 41% and 100% of cases.[10] Among guideline panel members, there was 'fair' inter-rater reliability for balance of benefit and harms (kappa = 0.4) and use of resources (kappa = 0.28), 'moderate' for patients' preferences and values (kappa = 0.44), 'fair' for assessing the strength of recommendations (kappa = 0.39), and 'good' for making recommendations (kappa = 0.74).[11]

The standard GRADE assessment had 'good' agreement (kappa = 0.66) with Trial Sequential Analysis for rating imprecision,[12] and 'fair' agreement (kappa = 0.35) for rating quality of evidence with the Semi-Automated Quality Assessment Tool, which is a 30-item checklist covering key determinants of the five GRADE domains.[13,14]

Commentary

GRADE is an essential tool for reviewers and decision-makers as it provides an indication of the confidence they can place in the results and a mechanism with which to translate the evidence into clinical practice guidelines. The initial work on GRADE reliability was conducted when limited guidance was available, resulting in many disagreements, but more detailed guidance seems to have improved inter-rater reliability. Specific training on GRADE methodology is recommended for inexperienced raters, and two independent raters are sufficient to reliably assess the quality of evidence. The basis for judgements should be made transparent and reported. Application of GRADE can be complex, as evidenced by a series of publications[15] and a lengthy handbook with over 10 chapters.[8] Different applications of GRADE and adaptations are likely to yield inconsistencies in ratings, which could influence decision-making. To maximise agreement, further research on assistive tools for GRADE assessment[13,14] is warranted.

Provenance: Invited. Not peer reviewed.

Charis X Xie and Gustavo C Machado
Institute for Musculoskeletal Health, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

References

1. Guyatt GH, et al. *BMJ*. 2008;336:924–926.
2. Atkins D, et al. *BMC Health Serv Res*. 2005;5:25.
3. Schünemann HJ, et al. *BMJ*. 2008;336:1106–1110.
4. Huguet A, et al. *Syst Rev*. 2013;2:71.
5. Lewin S, et al. *Implement Sci*. 2018;13:2.
6. Guyatt GH, et al. *BMJ*. 2008;336:1049–1051.
7. The GRADE Working Group. https://cebgrade.mcmaster.ca.
8. The GRADE Working Group. https://gradepro.org.
9. Mustafa RA, et al. *J Clin Epidemiol*. 2013;66:736–742.
10. Hartling L, et al. *PLoS One*. 2012;7:e34697.
11. Kumar A, et al. *J Clin Epidemiol*. 2016;75:115–118.
12. Castellini G, et al. *Syst Rev*. 2018;7:110.
13. Llewellyn A, et al. *PLoS One*. 2016;10:e0123511.
14. Meader N, et al. *Syst Rev*. 2014;3:82.
15. Guyatt GH, et al. *J Clin Epidemiol*. 2011;64:380–382.

https://doi.org/10.1016/j.jphys.2020.07.003