Trial sequential analysis for assessing imprecision in GRADE evaluations – protocol for a methodological study

Joachim Birch Milan[1] (ORCID: 0000-0001-7093-5432) / Johanne Periera Ribeiro[2,3] (ORCID: 0000-0001-6019-022X), Christian Gunge Riberholt[4] (ORCID: 0000-0002-6170-1869), Markus Harboe Olsen[1,5] (ORCID: 0000-0003-0981-0723), and Christian Gluud[1,6] (ORCID: 0000-0002-8861-0799)

Corresponding author: joachim.birch.milan@regionh.dk

[1] Copenhagen Trial Unit, Centre for Clinical Intervention Research, The Capital Region, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark
[2] Center for Evidence-Based Psychiatry, Psychiatric Research Unit, Psychiatry Region Zealand, Region Zealand, Denmark
[3] Department of Psychology, The Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark
[4] Department of Neurorehabilitation / Traumatic Brain Injury, Copenhagen University Hospital – Rigshospitalet, Glostrup, Denmark
[5] Department of Neuroanaesthesiology, The Neuroscience Centre, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark
[6] Department of Regional Health Research, The Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark

Abstract

Background: Assessing statistical imprecision of summary estimates is an essential element in evaluating the strength of evidence. The GRADE framework recommends assessing imprecision by confidence intervals (CI) in relation to thresholds of interest and in selected cases to assess the relationship between the acquired information size and the calculated optimal information size (OIS).

Trial sequential analysis (TSA) calculates TSA-adjusted confidence intervals after it has calculated a required information size. In a recent methodological study of 544 systematic reviews and meta-analysis reports of clinical trials with TSA, we gathered data regarding the methods used for grading imprecision, specifically regarding the impact of TSA (the METSA project).

The questions regarding GRADE imprecision were initially superficially defined and were substantialised only during the project and in the preparations for this protocol. With this add-on study, we investigate the methods of grading imprecision in the GRADE framework by authors of systematic reviews and meta-analysis reports of clinical trials with TSA.

Methods: The outlined methodological study will be pre-planned but designed with knowledge about existing but not yet reviewed data on the study questions. The METSA project was not initially designed for the questions raised in the current study protocol, which warrants a critical review of the collected data. We aim to improve precision and accuracy of the collected data regarding imprecision methodology by a redesign of selected data fields, adding new data fields to the data extraction form, and a subsequent revision of the existing data extraction accordingly. For each individual study, we will extract or review data regarding the specified methodology including methods for calculating CI and OIS, and thresholds of interest (definitions of important benefit and/or harm). For each topic, we will assess completeness in transparency of described methods and protocolisation, including coherence with the protocol (if relevant).

Results: We will report frequencies of observed methods, lack of transparency, and protocolisation. From data gathered in the METSA project, we will report the proportion of imprecision assessments that may have differed in their conclusions if the results of the TSA had been used. Informed by our findings, we will outline new suggestions on how to grade imprecision using TSA.

Conclusion: This protocol outlines a methodological study of methods and reporting characteristics imprecision assessment within the GRADE framework in recent systematic reviews and meta-analysis reports of clinical trials utilising TSA.

Keywords: Trial sequential analysis; imprecision; type I error; type 2 error; GRADE; systematic review; meta-analysis; research on research, reproducibility

Introduction

Conclusions from systematic reviews with meta-analysis depend on the confidence of the summary estimates (H Schünemann et al., 2013). The Cochrane Handbook encourages the use of The Grading of Recommendations, Assessment, Development, and Evaluations framework (GRADE) (HJ Schünemann, Higgins, et al., 2022) to evaluate the certainty of evidence in systematic reviews, clinical guidelines, population studies, and more (H Schünemann et al., 2013; HJ Schünemann, Higgins, et al., 2022; HJ Schünemann, Vist, et al., 2022).

Box 1

We define a *systematic review* (SR) as a protocolised approach to evidence synthesis, using verifiably pre-defined eligibility criteria and synthesis methodology.

We define a *meta-analysis report* (MAR) as methodologically comparable to a systematic review but lacking a predefined publicly available protocol.

In systematic reviews of randomised clinical trials the certainty of the evidence used to generate summary estimates for a given outcome is assessed with GRADE by five domains: (1) risk of bias, (2) publication bias, (3) imprecision, (4) indirectness, and (5) inconsistency (Balshem et al., 2011). In the GRADE framework, the certainty of the evidence is initially assumed high and subsequently downgraded to either moderate, low, or very low certainty, according to the five domains. Each domain can downgrade the certainty one, two, or – as recently suggested for imprecision – three levels (G H Guyatt, Oxman, Kunz, Brozek, et al., 2011; G H Guyatt, Oxman, Kunz, Woodcock, et al., 2011b, 2011a; G H Guyatt, Oxman, Montori, et al., 2011; Gordon H Guyatt et al., 2011; H Schünemann et al., 2013; Zeng et al., 2022).

Imprecision is a different term for statistical uncertainty and in this context primarily arises from underpowered meta-analyses, which are comparable to interim analyses of randomised clinical trials (Bender et al., 2008; Brok et al., 2009; Kjaergard et al., 2001). Assessment of imprecision is a key domain in the GRADE framework as recommendations about interventions cannot be made based on imprecise summary estimates.

In the METSA project (Riberholt et al., 2022), we assessed 544 systematic reviews and meta-analysis reports (see Box 1) regarding the use of trial sequential analysis (TSA). The focus of the METSA project was assessing transparency and completeness in reporting of TSA and the related conclusions, e.g. assessing imprecision in the GRADE framework.

In the outlined methodological study, we focus on the methodology and reporting characteristics of imprecision assessments in the GRADE framework in systematic reviews and meta-analysis reports with TSA with the purpose of contributing to the further development and promotion of the GRADE framework.

Methods

Study design

This protocol outlines a pre-planned, secondary report on the current practice of grading imprecision within the GRADE framework in systematic reviews and meta-analysis reports of clinical trials, which utilised TSA. The aim of the outlined study was defined in the METSA project protocol (Riberholt et al., 2022), but the methods applied are defined post-hoc.

Data material

We will adapt the existing METSA project database. The METSA project database, which is available at zenodo.org (DOI: 10.5281/zenodo.8318331), contains data on 544 systematic reviews and meta-analysis reports, 300 (55%) of which applied GRADE. A complete description of the METSA project methodology is available at (Riberholt et al., 2022) [REF also? to main paperSee my comment above]. In brief, we searched MEDLINE and the Cochrane database for systematic reviews and meta-analysis reports of randomised clinical trials which utilised TSA published between January 2018 and January 2022. For each included study, we extracted baseline data and assessed the study using AMSTAR 2 (Shea Beverley et al., 2017). We extracted data regarding TSA on one dichotomous outcome analysis (n=439) and one continuous outcome analysis (n=185), if applicable (total TSAs n = 624). All tasks regarding literature search, data extraction, and AMSTAR assessment were performed in duplicate by study authors using predefined criteria in a standardised data extraction form.

The METSA project was not initially designed for the questions raised in the current study protocol, and upon data revision, we have found that a critical review of the collected data is warranted.

Data extraction

All tasks regarding data revision and extraction will be performed in duplicate by two independent authors. Discrepancies will be resolved through a consensus process. A third author will be involved in case of disagreements.

We aim to improve precision and accuracy of the collected data regarding imprecision methodology by a redesign of selected data fields, adding new data fields to the data extraction form and a subsequent revision of the existing data extraction accordingly. The amended data extraction form is provided in Supplement. Most changes concern specifications regarding the specified methodology including methods for calculating CI and RIS, and thresholds of interest (definitions of important benefit and/or harm). For each topic, we will assess completeness in transparency of described methods and protocolisation, including coherence with the protocol (if relevant).

Analysis and presentation of results

We will report and describe the observed methods for downgrading imprecision, lack of transparency, and protocolisation and report the frequencies of each. The data extraction form has predefined options for certain fields but not for other. We will assess the need for revising the defined options/categories based on the collected data. Specifically, we will report:

- Frequencies of publications with each identified method for downgrading imprecision, including frequencies of unclear methodology
    - For relevant methods, the frequency of setting limits for important differences for benefit and harm, respectively
    - For relevant methods, the frequency of reporting OIS calculation methods (if applicable, we will provide an overview of identified methods)
- Frequencies of publications in which the specified methodology:
    - Was not planned in a protocol made public prior to data extraction
    - Differed from the methodology described in the protocol
- Frequencies of outcomes downgraded 0, 1, 2, or 3 levels respectively, including frequencies of unclear or non-reproducible gradings of imprecision.

From data gathered in the METSA project, we will additionally report the proportion of imprecision assessments that may have differed if they had assessed imprecision using the results of their TSA analysis, i.e. had used TSA CI instead of conventional 95% CI (only regarding inclusion of no effect, i.e. not important benefits or harms) or an RIS instead of OIS for downgrading imprecision. We will seek to demonstrate how TSA may contribute to downgrading imprecision with practical examples from the included studies in relation to a discussion of the statistical theory behind TSA. Informed by our findings, we will develop suggestions on how to assess imprecision in the GRADE framework using TSA, including reporting standards. We will discuss these new suggestions in relation to previous suggestions of using TSA for GRADE.

Deviations from the outlined analysis and presentation plan will be described in the final study report.

Discussion

In this protocol, we outline a methodological study in which we will assess and report the methodology of downgrading imprecision in the GRADE framework in recent systematic reviews and meta-analysis reports of randomised clinical trials which use TSA. The aim of the study is to bring attention to the important topic of imprecision by providing an overview of the methods used in current literature and potentially identifying inadequacies in methodology and reporting. It is our goal to contribute to the further development and promotion of the GRADE framework.

This study is not without limitations. Three study authors (JBM, JMPR, CGR) participated in extracting, revising, and reviewing data for the METSA project. The answers to the questions raised in this protocol are partially known by these authors, e.g. there is a significant lack of transparency in the utilised methods, but with uncertainty due insufficient data accuracy and precision. These observations have not been shared elsewhere. The outlined study seeks to answer new questions as well as confirming the assumptions that were based on observations made during the METSA project.

Conclusion

This protocol outlines a methodological study of method and reporting of imprecision assessment within the GRADE framework in recent systematic reviews and meta-analysis reports of clinical trials utilising trial sequential analysis.

Additional information

Project status

Data revision regarding the variables relevant for the outlined study has been initiated (n=25) for the purpose of testing the data extraction form.

Ethical considerations

The outlined study is performed on public, non-sensitive data.

Author contributions

JPR and JBM are responsible for the design of the outlined study and drafted the protocol manuscript.

All authors contributed to the design of the outlined study, and critically revised and approved the final version of the protocol manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

For contributions to the METSA database, we refer to the METSA publication (awaiting publication).

Sources of funding and conflicts of interest

Neither the outlined study nor the METSA project received external financial support. Java and R implementations of trial sequential analysis was developed by the Copenhagen Trial Unit, for R directed by Anne Lyngholm Sørensen (https://orcid.org/0000-0002-8265-0394), PhD student at Section for Biostatistics, Institute of Public Health, Copenhagen University. The authors have nothing to declare.

Data and source code availability

The METSA project database is available at zenodo.org (DOI: 10.5281/zenodo.8318331). The amended METSA project database containing data used for the outlined study will be made available along with any relevant code for analysis at zenodo.org.

Acknowledgements

We are grateful to Mark Asante (https://orcid.org/0009-0002-8034-4139) and Buddheera Kumburegama for contributing to the data revision for this study.

References

Balshem, H., Helfand, M., Schünemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., Vist, G. E., Falck-Ytter, Y., Meerpohl, J., Norris, S., & Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology*, *64*(4), 401–406. https://doi.org/10.1016/j.jclinepi.2010.07.015

Bender, R., Bunce, C., Clarke, M., Gates, S., Lange, S., Pace, N. L., & Thorlund, K. (2008). Attention should be given to multiplicity issues in systematic reviews. *Journal of Clinical Epidemiology*, *61*(9), 857–865. https://doi.org/10.1016/J.JCLINEPI.2008.03.004

Brok, J., Thorlund, K., Wetterslev, J., & Gluud, C. (2009). Apparently conclusive meta-analyses may be inconclusive - Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. *International Journal of Epidemiology*, *38*(1), 287–298. https://doi.org/10.1093/ije/dyn188

Guyatt, G H, Oxman, A. D., Kunz, R., Brozek, J., Alonso-Coello, P., Rind, D., & Et, a l. (2011). GRADE guidelines: 6. Rating the quality of evidence - imprecision. *Journal of Clinical Epidemiology*, *64*(12), 1283–93–1283–93. https://doi.org/10.1016/j.jclinepi.2011.01.012

Guyatt, G H, Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., & et, a l. (2011a). GRADE guidelines: 7. Rating the quality of evidence - inconsistency. *Journal of Clinical Epidemiology*, *64*(12), 1294–302–1294–302.

Guyatt, G H, Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., & et, a l. (2011b). GRADE guidelines: 8. Rating the quality of evidence - indirectness. *Journal of Clinical Epidemiology*, *64*(12), 1303–10–1303–10.

Guyatt, G H, Oxman, A. D., Montori, V., Vist, G., Kunz, R., Brozek, J., & et, a l. (2011). GRADE guidelines: 5. Rating the quality of evidence - publication bias. *Journal of Clinical Epidemiology*, *64*(12), 1277–82–1277–82.

Guyatt, Gordon H, Oxman, A. D., Vist, G., Kunz, R., Brozek, J., Alonso-Coello, P., Montori, V., Akl, E. A., Djulbegovic, B., & Falck-Ytter, Y. (2011). GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). *Journal of Clinical Epidemiology*, *64*(4), 407–415. https://doi.org/10.1016/j.jclinepi.2010.07.017

Kjaergard, L. L., Villumsen, J., & Gluud, C. (2001). Reported Methodologic Quality and Discrepancies between Large and Small Randomized Trials in Meta-Analyses. *Annals of Internal Medicine*, *135*(11), 982–989. https://doi.org/10.7326/0003-4819-135-11-200112040-00010

Riberholt, C. G., Olsen, M. H., Milan, J. B., & Gluud, C. (2022). Major mistakes and errors in the use of Trial Sequential Analysis in systematic reviews or meta-analyses – protocol for a systematic review. *Systematic Reviews*, *11*(1), 114. https://doi.org/10.1186/s13643-022-01987-4

Schünemann, H, Brożek, J., Guyatt, G., & Oxman, A. (2013). *GRADE handbook for grading quality of evidence and strength of recommendations. Available from guidelinedevelopment.org/handbook* (Holger Schünemann, J. Brożek, G. Guyatt, & A. Oxman (eds.); Updated Oc). The GRADE Working Group.

Schünemann, HJ, Higgins, J., Vist, G., Glasziou, P., Akl, E., Skoetz, N., & Guyatt, G. (2022). Chapter 14:

Completing 'Summary of findings' tables and grading the certainty of the evidence. In J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane.

Schünemann, HJ, Vist, G., Higgins, J., Santesso, N., Deeks, J., Glasziou, P., Akl, E., & Guyatt, G. (2022). Chapter 15: Interpreting results and drawing conclusions. In J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, & V. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane.

Shea Beverley, J., Reeves Barnaby, C., Wells, G. e. o. r. g. e., Thuku, M. i. c. e. r. e., Hamel, C. a. n. d. y. c. e., Moran, J. u. l. i. a. n., Moher, D. a. v. i. d., Tugwell, P. e. t. e. r., Welch, V. i. v. i. a. n., & Kristjansson, E. l. i. z. a. b. e. t. h. (2017). AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *Bmj*, *358*, j4008–j4008.

Zeng, L., Brignardello-Petersen, R., Hultcrantz, M., Mustafa, R. A., Murad, M. H., Iorio, A., Traversy, G., Akl, E. A., Mayer, M., Schünemann, H. J., & Guyatt, G. H. (2022). GRADE Guidance 34: update on rating imprecision using a minimally contextualized approach. *Journal of Clinical Epidemiology*, *150*(20), 216–224. https://doi.org/10.1016/j.jclinepi.2022.07.014