25 September 2023

Reliability, usability, and coverage of AMSTAR 2 assessing 544 systematic reviews and meta-analysis reports – protocol for a descriptive analytic study

Joachim Birch Milan[1] (ORCID: 0000-0001-7093-5432), Christian Gunge Riberholt[2] (ORCID: 0000-0002-6170-1869), Markus Harboe Olsen[1,3] (ORCID: 0000-0003-0981-0723), Zheng Yii Lee[4,5] (ORCID: 0000-0003-4505-7476), Johanne Pereira Ribeiro[6,7] (ORCID: 0000-0001-6019-022X), Charles Chin Han Lew[8] (ORCID: 0000-0001-6410-3859), and Christian Gluud[1,9] (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] Department of Neurorehabilitation / Traumatic Brain Injury, Copenhagen University Hospital - Rigshospitalet, Glostrup, Denmark

[3] Department of Neuroanaesthesiology, The Neuroscience Centre, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark

[4] Department of Anaesthesiology, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia.

[5] Department of Cardiac Anesthesiology & Intensive Care Medicine, Charité Berlin, Germany

[6] Center for Evidence-Based Psychiatry, Psychiatric Research Unit, Psychiatry Region Zealand, Region Zealand, Denmark

[7] Department of Psychology, The Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark

[8] Department of Dietetics & Nutrition, Ng Teng Fong General Hospital, Singapore

[9] Department of Regional Health Research, The Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark

25 September 2023

Abstract

Background: The AMSTAR 2 tool (second version of Assessing Methodological Quality in Systematic Reviews) is useful for critical appraisal of systematic reviews of clinical trials. In a methodological study of systematic reviews and meta-analysis reports of randomised clinical trials which used the sequential meta-analysis trial sequential analysis (the METSA study), we used the AMSTAR 2 to assess the overall quality of each included study. With the study outlined in this protocol, we aim to explore the interrater reliability of the AMSTAR 2, qualitatively describe our experiences using the tool, and discuss the tool's coverage of critical domains.

Methods: In the METSA study, we investigated statistical methodology and transparency in 544 systematic reviews and meta-analysis reports of randomised clinical trials which used trial sequential analysis (TSA). All systematic reviews (with a protocol) were assessed with AMSTAR 2 by two independent authors (n=270). Meta-analysis reports – defined as not having a protocol – were automatically rated as 'critically low confidence' and did no undergo further AMSTAR 2 assessment. Disagreement on the AMSTAR 2 rating was resolved through discussion between the authors. Principal issues were discussed at weekly meetings. Thoughts on the usability and coverage of AMSTAR 2 was shared at these meetings and noted throughout and will be collected post-hoc for the current study. Here, we will analyse the level of agreement on the initial ratings by raw agreement rates and Cohen's kappa and test for trends concerning the effect of the consensus process (rating up or down confidence) as well as the overall effect of assessor experience. We will compare the AMSTAR 2 rating with the assessments of TSA transparency performed during the METSA study.

Conclusion: This methodological study will provide insights in some of the characteristics of AMSTAR 2, including interrater reliability and usability in the context of assessing 270 systematic reviews of clinical trials. We will provide group consensus-based suggestions regarding usability and coverage.

Keywords: systematic review, AMSTAR 2, meta-analysis, evidence-based medicine

Introduction

Systematic reviews (SR) of randomised clinical trials are generally considered the highest level of evidence in clinical science (Garattini et al., 2016). The validity of SRs hinges on the methodological robustness of the SR. Methodological issues in SRs prevail and therefore, a thorough, valid, and systematic approach to critical appraisal of SRs is essential for evidence-based medicine (Garattini et al., 2016). AMSTAR (A MeaSurement Tool to Assess systematic Reviews) represents such a systematic approach and has become a popular tool in addressing issues of individual systematic reviews (De Santis et al., 2023; Shea et al., 2007).

The revised AMSTAR tool (AMSTAR 2) was published in 2017 to increase the number of critical domains covered and to be more user friendly with easier response categories and better guidance (Shea et al., 2017). However, AMSTAR 2 continues to be inappropriately applied, suggesting a need for even clearer guidance on AMSTAR assessment and reporting (Pieper et al., 2018).

In a methodological study of 544 systematic reviews (with a verifiably pre-planned protocol) and meta-analysis reports (without a verifiably pre-planned protocol) of clinical trials which applied trial sequential analysis (the METSA study), we investigated statistical methodology and transparent reporting of trial sequential analysis (Riberholt et al., 2022). Trial sequential analysis (TSA) is a meta-analysis method based on Lan-DeMets alpha spending boundaries that controls the risk of false positives due to repeated testing (a concept best known from interim analyses in single trials) (Wetterslev et al., 2017). We used the AMSTAR 2 to assess the overall quality of each included SR. In the outlined study, we will share our experience of using the AMSTAR 2 for critical appraisal of SR in the METSA study, including assessment of reliability, usability, and coverage, and hope to contribute to the further development of the AMSTAR system.

Methods

This protocol outlines a post-hoc descriptive analytic study of the AMSTAR 2 reliability, usability, and coverage. The aim of the outlined study was not defined in the METSA project protocol, and the methods applied are defined post-hoc of the METSA project (Riberholt et al., 2022).

Data material

The AMSTAR 2 assessment was performed as part of the METSA study, which is a methodological study of 544 systematic reviews and meta-analysis reports of clinical trials using trial sequential analysis (Riberholt et al., 2022).

In brief, we searched MEDLINE and the Cochrane Database for Systematic Reviews for SR and meta-analysis reports of clinical trials which utilised trial sequential analysis published between January 2018 and January 2022. For each included study, we extracted characterising data (country of publication, population, intervention, comparator, and outcomes, number of trials included, etc.) and assessed the study using AMSTAR 2. For each study, we extracted data regarding TSA on one dichotomous outcome analysis (n=439) and one continuous outcome analysis (n=185), if applicable (total n = 624). All tasks regarding literature search, data extraction, and AMSTAR 2 assessment were performed in duplicate by study authors using predefined criteria in a standardized data extraction form.

Method of AMSTAR 2 assessment

AMSTAR 2 was incorporated in our standardised data extraction form in REDCap (Research Electronic Data Capture) (Harris et al., 2009) for this project.

Each included study was assessed independently by two authors from the assessor group. The authors assigned themselves for study assessment on an ad-hoc basis. After completed data extraction for each included study, the two authors sought consensus on the final rating through discussion. Persistent disagreements or principal issues were discussed and resolved at weekly research meetings. AMSTAR 2 assessment was always performed before any other data extraction, to minimise the impact of the latter on the former, although consensus was sought only after completed data extraction.

Included studies that did not have a documentable pre-defined protocol were all considered meta-analysis reports (MAR) and were rated as of 'critically low confidence' (274/544). The individual AMSTAR 2 items were not assessed further, as further assessment would not impact the overall rating. Therefore, these studies will not be included for the current analysis. Studies with a documentable pre-defined protocol were considered systematic reviews (SR) and were all assessed using each of the 16 items in AMSTAR 2.

We did not modify the AMSTAR 2 tool, however, in our data extraction form, we added an automated calculator for each item concluding 'Yes', 'Partially yes', or 'No', corresponding to the original AMSTAR 2 tool. To each item, we further added a multiple-choice field, e.g., 'Did the "PICO" question reveal any moderate or critical weaknesses?', with the answer options 'Yes, critical weakness', 'Yes, moderate weakness', and 'No'. We further added an optional comment field under each item, where weaknesses noted for each item could be described. The presence of critical flaws and non-critical

weaknesses were listed in an auto-generated table at the bottom of the form for easy overview. The number of critical flaws and the occurrence of 'Yes', 'Partially yes' and 'No' were not calculated or in other ways analysed, and the rating relied on an overall assessment, as is recommended by the AMSTAR 2 guidance document.

The assessors

The AMSTAR 2 assessor group consisted of the 13 data extractors from the METSA study. At the time of data extraction and consensus, four assessors were medical students, one was a medical doctor, two were clinical dietitians with PhDs, two were Masters of Public Health in PhD-programmes, two were physiotherapists with PhDs, one was physiotherapist in a PhD-programme, one was a psychologist in a PhD-programme, and one was a medical doctor in a PhD-programme. Some assessors were familiar with the AMSTAR 2 tool, however, no one in the assessor group had formal experience with applying the tool, except CGR, CG, and JPR. CG and CGR instructed the assessors in use of AMSTAR 2 prior to initiation of each assessor's participation in the data extraction process.

Statistical analysis plan

Data from the METSA database in REDCap will be exported and analysed in the latest available stable version of R (R Core Team, 2022). Meta-analysis reports (without a pre-published protocol) are excluded from the analysis.

Interrater reliability will be analysed by calculating raw agreement rates for each AMSTAR level (after consensus) and Cohen's kappa and weighted kappa coefficients. We will further calculate raw agreement rates for each variable in the AMSTAR assessment (each checkable answer option, the calculated AMSTAR 'conclusion' for each item; 'Yes', 'Partially yes', or 'No', and the appended question of whether weaknesses were identified for each item).

To assess whether the individual assessors had distinct rating tendencies, e.g. more positive, or negative, we will compare each reviewer's initial rating with the corresponding rating of each study. An initial rating that was identical to the corresponding rating will receive a score of 0, while an assessment that was more positive, e.g. 'Moderate' against 'Low' will receive a score of +1, 'High' against 'Critically low' will receive +3, 'Low' against 'Critically low' will receive -1, etc. We will then calculate the mean for each assessor. This method may be biased as assessors may have tended to co-assess with, e.g. other positive assessors. To partially account for this, we will provide an overview of co-assessments in a network graph.

A mixed effects ordinal regression will be used to ascertain if the overall rating tendency is influenced by experience gained over time, by using the rating as an ordinal outcome, assessor as random effects, and accumulated number of systematic reviews assessed as a fixed effect.

We will analyse the change in rating after consensus was performed, to see if the consensus process generally made the AMSTAR ratings more positive or more negative. For each individual study, we will calculate the sum of difference (e.g. two initial ratings at 'High' and 'Low' which after consensus are changed to 'Moderate' have a change of -1 and +1, respectively, with a sum of 0. If the rating had been changed into 'Low', the changes would be -2 and 0, respectively, summing to -2) and

provide a table for the frequency of each possible sum of difference (–5 to +5) subgrouped by difference between initial ratings (1 to 3).

To test whether AMSTAR 2 is potentially insensitive to transparent reporting of statistical methods, we will test the correlation between the AMSTAR rating and the TSA transparency ratings of each study by ordinal regression. If possible, we will test the correlation between AMSTAR rating and the secondarily collected GRADE imprecision transparency assessments of each study (protocol: 10.5281/zenodo.8318950).

For all frequentist analyses, we will not perform null hypothesis significance testing, but will calculate 95% confidence intervals where relevant.

Qualitative evaluation of AMSTAR 2 usability and coverage

All assessors were on multiple occasions encouraged to take note of issues or challenges regarding usability or coverage of the AMSTAR 2 assessments and report these at the weekly meetings or in a shared project document. We will read all comments in the comment fields that we added to the AMSTAR 2 segment of our data extraction form to identify comments indicating assessment issues or challenges. All assessors will be requested to provide feedback on coverage and usability of the AMSTAR 2 tool, either in written or oral communication.

Reporting of results

We will report and discuss the results of all conducted statistical analyses as defined in the statistical analysis plan.

We will report raw agreement rates for each variable in the AMSTAR assessment and the overall rating, and Cohen's kappa for the overall rating. We will report the frequencies of disagreement levels (0 levels = no disagreement, 1 level = minor disagreement, 2 levels = major disagreement, 3 levels = extreme disagreement).

For each reviewer, we will provide the mean difference between the reviewers initial rating and the corresponding rating. We will provide a visual overview of how the individual assessors teamed up in a network graph.

We will provide a table for the frequency of each possible sum of differences between initial ratings and consensus rating (-5 to +5) subgrouped by difference between initial ratings (1-3).

For the ordinal regression models, we will report odds ratios (with 95% CI) and measures of goodness-of-fit. We will provide a plot of each assessor's ratings in a chronological order.

We will further provide a qualitative description of the feedback provided during the assessment process as well as feedback received after the initiation of the outlined study. We will also report if any comments made directly in the AMSTAR assessment indicate an issue with usability or coverage.

Discussion

The AMSTAR 2 tool is generally considered a useful, valid, and reliable tool for critical appraisal of systematic reviews of randomised clinical trials. However, previous reports by AMSTAR 2 users suggest a need for improved usability and guidance (Pieper et al., 2018). Additionally, AMSTAR 2 has been critiqued for being superficial in the description of included domains, e.g. conflicts of interest (Lundh et al., 2020), lacking clear reasoning behind the definitions of critical domains (Li et al., 2022) and additionally lacking guidance on some domains (De Santis et al., 2023). In our outlined study, we will provide a detailed discussion of our assessor groups opinions on the usability and coverage, e.g. a discussion of the AMSTAR 2 instrument in relation to assessment of trial sequential analysis (the focus of the METSA project) and the GRADE guidelines (Schünemann et al., 2013).

Systematic review methodology is a field in constant development and so, continuous updating and improvement of the AMSTAR tool is warranted. The outlined study aims to contribute with insights into the further development of the AMSTAR tool.

Interpretation of results

We will seek to identify outliers in the overview of agreement rates for each variable, as these may indicate particularly challenging items or domains. From the agreement rates and kappa scores for the overall rating, we will discuss the reliability of the AMSTAR 2 tool.

From the analyses of effects of experience and distinct rating tendencies per reviewer, we will discuss the potential impact of assessor selection and the relative importance of training and experience before using AMSTAR 2 for critical tasks, such as guideline formations. The findings of these analyses will potentially be biased, particularly if we see that the assessors tended to co-assess with other specific assessors.

With the table of frequencies of each possible sum of difference (-5 to +5) subgrouped by difference between initial ratings (1-3), we will be able to identify common patterns in the effect of consensus process, e.g. if large disagreements (e.g. 'High' and 'Critically low') commonly result in a 'compromise' (e.g. 'Low', which would sum to -1) or choosing either initial rating (e.g., 'High', which would sum to +3 or 'Critically low', which would sum to -3). We may also observe that the consensus rating is sometimes lower than either of the initial ratings (e.g. initial ratings of 'Moderate' and 'Low' with consensus rating 'Critically low', summing to -3).

If we find that there is no correlation between the AMSTAR rating and the TSA transparency rating in each study, this could indicate that the AMSTAR tool puts too little emphasis on detailed transparent reporting.

Limitations

During the project, the assessor group developed a discourse on AMSTAR assessment through the consensus processes and the research meetings, which is expected to increase interrater agreement rates. Therefore, our findings may not be applicable to agreement between naïve assessors or assessors having a different group discourse.

In the case of AMSTAR assessment, interrater agreement is not directly tied to the tool's validity. The domains covered by AMSTAR 2 are complex and each individual researcher can validly hold differing opinions on whether a methodological choice is a critical flaw or not.

Conclusion

In this protocol, we describe a planned methodological study that will quantitatively and qualitatively assess reliability, usability, and coverage of the AMSTAR 2 assessment tool in the context of assessing 544 systematic reviews and meta-analysis reports of clinical trials. The study results will provide a basis for possibly making suggestions to recommended amendments of the AMSTAR 2, contributing to the further development.

Additional information

Project status

None of the data regarding AMSTAR in METSA has been viewed or analysed, except for the results provided in this protocol (proportion of studies at each AMSTAR level).

Ethical considerations

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

Author contributions

JBM, CGR, MHO and CG are responsible for study conception and design.JBM drafted the protocol manuscript.

All authors critically revised and approved the final version. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Sources of funding and conflicts of interest

Neither the outlined study nor the METSA project received external financial support. 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 source code used for the outlined study will be made available at zenodo.org.

Acknowledgements

None.

References

De Santis, K. K., Pieper, D., Lorenz, R. C., Wegewitz, U., Siemens, W., & Matthias, K. (2023). User experience of applying AMSTAR 2 to appraise systematic reviews of healthcare interventions: a commentary. *BMC Medical Research Methodology*, *23*(1), 63. https://doi.org/10.1186/s12874-023-01879-8

Garattini, S., Jakobsen, J. C., Wetterslev, J., Bertelé, V., Banzi, R., Rath, A., Neugebauer, E. A. M., Laville, M., Masson, Y., Hivert, V., Eikermann, M., Aydin, B., Ngwabyt, S., Martinho, C., Gerardi, C., Szmigielski, C. A., Demotes-Mainard, J., & Gluud, C. (2016). Evidence-based clinical practice: Overview of threats to the validity of evidence and how to minimise them. *European Journal of Internal Medicine*, *32*, 13–21. https://doi.org/10.1016/j.ejim.2016.03.020

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Li, L., Asemota, I., Liu, B., Gomez-Valencia, J., Lin, L., Arif, A. W., Siddiqi, T. J., & Usman, M. S. (2022). AMSTAR 2 appraisal of systematic reviews and meta-analyses in the field of heart failure from high-impact journals. *Systematic Reviews*, *11*(1), 147. https://doi.org/10.1186/s13643-022-02029-9

Lundh, A., Rasmussen, K., Østengaard, L., Boutron, I., Stewart, L. A., & Hróbjartsson, A. (2020). Systematic review finds that appraisal tools for medical research studies address conflicts of interest superficially. *Journal of Clinical Epidemiology*, *120*, 104–115. https://doi.org/10.1016/j.jclinepi.2019.12.005

Pieper, D., Koensgen, N., Breuing, J., Ge, L., & Wegewitz, U. (2018). How is AMSTAR applied by authors – a call for better reporting. *BMC Medical Research Methodology*, *18*(1), 56. https://doi.org/10.1186/s12874-018-0520-z

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

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. J., Brozek, J., Guyatt, G. H., & Oxman, A. D. (2013). *Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. Updated October 2013.* https://gdt.gradepro.org/app/handbook/handbook.html

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., Porter, A. C., Tugwell, P., Moher, D., & Bouter, L. M. (2007). Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, *7*(1), 10. https://doi.org/10.1186/1471-2288-7-10

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (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. https://doi.org/10.1136/bmj.j4008

Wetterslev, J., Jakobsen, J. C., & Gluud, C. (2017). Trial Sequential Analysis in systematic reviews with meta-analysis. *BMC Medical Research Methodology*, *17*(1), 1–18. https://doi.org/10.1186/s12874-017-0315-7