Journal of Clinical Epidemiology 125 (2020) 9−15

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

# Quality and risk of bias appraisals of systematic reviews are inconsistent across reviewers and centers

Michelle Gates[a], Allison Gates[a], Gonçalo Duarte[b], Maria Cary[c], Monika Becker[d], Barbara Prediger[d], Ben Vandermeer[a], Ricardo M. Fernandes[b,e], Dawid Pieper[d], Lisa Hartling[a,*]

[a]*Department of Pediatrics, Alberta Research Centre for Health Evidence, University of Alberta, Edmonton, Alberta, Canada*
[b]*Clinical Pharmacology Unit, Instituto de Medicina Molecular, University of Lisbon, Lisbon, Portugal*
[c]*Centre for Health Evaluation & Research (CEFAR), National Association of Pharmacies, Lisbon, Portugal*
[d]*Department für Humanmedizin, Institut für Forschung in der Operativen Medizin, Universität Witten/Herdecke, Witten, Germany*
[e]*Department of Pediatrics, Santa Maria Hospital, Lisbon, Portugal*

Accepted 16 April 2020; Published online 19 May 2020

## Abstract

**Objective:** The objective of the study was to evaluate the inter-rater and intercenter reliability, usability, and utility of A MeaSurement Tool to Assess systematic Reviews (AMSTAR), AMSTAR 2, and Risk Of Bias In Systematic reviews (ROBIS).

**Study Design and Setting:** This is a prospective evaluation using 30 systematic reviews of randomized trials, undertaken at three international centers.

**Results:** Reviewers completed AMSTAR, AMSTAR 2, and ROBIS in median (interquartile range) 15.7 (11.3), 19.7 (12.1), and 28.7 (17.4) minutes and reached consensus in 2.6 (3.2), 4.6 (5.3), and 10.9 (10.8) minutes, respectively. Across all centers, inter-rater reliability was substantial to almost perfect for 8/11 AMSTAR, 9/16 AMSTAR 2, and 12/24 ROBIS items. Intercenter reliability was substantial to almost perfect for 6/11 AMSTAR, 12/16 AMSTAR 2, and 7/24 ROBIS items. Intercenter reliability for confidence in the results of the review or overall risk of bias was moderate (Gwet's first-order agreement coefficient (AC1) 0.58, 95% confidence intervals [CI]: 0.30 to 0.85) to substantial (AC1 0.74, 95% CI: 0.30 to 0.85) for AMSTAR 2 and poor (AC1 −0.21, 95% CI: −0.55 to 0.13) to moderate (AC1 0.56, 95% CI: 0.30 to 0.83) for ROBIS. It is not clear whether using the appraisals of any tool as an inclusion criterion would alter an overview's findings.

**Conclusions:** Improved guidance may be needed to facilitate the consistent interpretation and application of the newer tools (especially ROBIS). 

*Keywords:* Systematic reviews; AMSTAR; AMSTAR 2; ROBIS; Methodological quality; Risk of bias

## 1. Introduction

Overviews of reviews (overviews) integrate evidence from multiple systematic reviews (SRs) and have become an increasingly popular [1−3] mechanism to help knowledge users cope with the growing body of SRs. The value of overviews depends on the quality of their included SRs. However, methodological and reporting shortcomings are common among SRs in the biomedical sciences [4].

Until recently, A MeaSurement Tool to Assess systematic Reviews (AMSTAR) [5] was the most commonly recommended tool to appraise the methodological quality of SRs of randomized controlled trials (RCTs) [6]. The tool contains 11 items with 4 response categories (yes, no, cannot answer, and not applicable), but no 'overall' score

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Dr. Lisa Hartling is supported by a Tier 1 Canada Research Chair in Knowledge Synthesis and Translation. The funding body had no role in the study design; collection, analysis, and interpretation of data; writing the report; or in the decision to submit the article for publication.

Data statement: The data pertaining to this study are available from the authors on reasonable request.

Conflict of interest: L.H. is supported by a Tier 1 Canada Research Chair in Knowledge Synthesis and Translation. Other authors have no potential conflicts of interest to declare.

* Corresponding author. 4-472 Edmonton Clinic Health Academy, 11405-87 Ave NW, Edmonton, Alberta, T6G 1C9, Canada. Tel.: +1-780-492-6124.

*E-mail address:* hartling@ualberta.ca (L. Hartling).

https://doi.org/10.1016/j.jclinepi.2020.04.026

**What is new?**

**Key findings**

- Compared to A MeaSurement Tool to Assess systematic Reviews (AMSTAR), the application of AMSTAR 2 takes slightly longer and Risk Of Bias In Systematic reviews (ROBIS) substantially longer.
- Inter-rater reliability is highly variable for AMSTAR 2 and even more so for ROBIS. Intercenter agreement is lower and more variable for ROBIS than the two AMSTAR tools.
- It is unclear whether using the ratings of AMSTAR, AMSTAR 2, or ROBIS as an inclusion criterion for systematic reviews would alter an overview of review's results or conclusions.

**What this adds to what was known?**

- This is the first study to directly compare AMSTAR, AMSTAR 2, and ROBIS for inter-rater and inter-centre reliability.
- This was the first study to explore the utility of the tools in informing the inclusion of SRs in overviews.

**What is the implication and what should change now?**

- Authors of overviews of reviews should report their decision rules to facilitate readers' interpretation of risk of bias or quality ratings.
- When possible, available guidance should be explicit, especially when multiple response options exist, to avoid misinterpretation and to encourage the consistent application of AMSTAR 2 and, even more so, ROBIS.

[5]. After further development to allow for the appraisal of SRs of nonrandomized studies, AMSTAR 2 was launched [7]. The tool contains 16 items with simplified response categories (yes, partial yes, and no) and allows reviewers to assign an overall confidence rating to the SR [7]. Risk Of Bias In Systematic reviews (ROBIS) is the first tool to facilitate the appraisal of risk of bias (ROB) in SRs [8]. Reviewers identify study limitations across four key domains using signaling questions, each with up to five response options (yes, probably yes, no information, probably no, and no) [8]. Reviewers may then judge the overall ROB in the review [8].

All three tools exhibit good face and content validity [5,7,8], and the inter-rater reliability (IRR) for most AMSTAR items is substantial [9]. The developers of AMSTAR 2 reported at least moderate IRR for most items [7]. Others have found slightly lower IRR for several items, and IRR for overall confidence in SR findings has ranged from fair [10] to moderate [11]. Reported IRR for ROBIS domains has varied widely, ranging from fair to substantial [10−15]. Reliability of ratings between pairs of reviewers at independent evidence review centers remains unknown. Previous reports have estimated that completing AMSTAR takes 10 to 20 minutes [7,9,16], AMSTAR 2 takes 15 to 32 minutes [7,10] per SR, and ROBIS takes markedly longer [12,14]. There is limited evidence on the utility of the tools to reviewers (i.e., if the assessments may be used to include or exclude SRs from overviews without changing the overview's results and/or conclusions) of the three tools to informing the inclusion of SRs in overviews, and to our knowledge, this is the first study to compare all three tools directly.

Given the recent introduction of AMSTAR 2 and ROBIS, evidence of their relative reliability, usability, and utility is needed. The current lack of empiric evidence available to understand how AMSTAR 2 or ROBIS compare with AMSTAR and to make recommendations for one tool over another [17] made it important to also include the older AMSTAR tool in our evaluation. Using a sample of SRs of health-care interventions, for each tool, we determined the following: (a) IRR for pairs of reviewers, and for the consensus of reviewer pairs at three evidence synthesis centers, (b) usability, based on the time to complete the appraisals and reach consensus, and (c) utility, based on whether their findings could inform the inclusion of SRs in overviews without changing the overview's results and/or conclusions.

## 2. Methods

Detailed methods for this prospective evaluation are in the published protocol [18] and are briefly described in sections 2.1 to 2.4. The study was undertaken by three international evidence synthesis centers in Canada, Germany, and Portugal.

### 2.1. Sample selection

Based on convenience and resource constraints, we chose to use a descriptive analysis of SRs published by Page et al. [4] as the source of reviews for our test sample. The analysis by Page et al. included a random sample of 300 SRs of biomedical and public health research indexed in MEDLINE in February 2014 [4]. From these, we randomly selected 30 SRs of RCTs of therapeutic interventions. For clarity and simplicity, we chose to include only SRs of RCTs (and not SRs of nonrandomized studies), which facilitated comparisons to AMSTAR (which is not intended for the appraisal of SRs of nonrandomized

studies) and limited the potential that our observations would be confounded by the type of SR evaluated.

### 2.2. Data collection

All data were collected following a predetermined data extraction guide and stored in Excel (v. 2016, Microsoft Corporation, Redmond, WA) worksheets.

#### 2.2.1. Characteristics of the reviews

From each SR, one reviewer extracted the publication and participant characteristics and eligible interventions and comparators. Another reviewer verified the extraction. Two reviewers independently determined the primary outcome of the SR, and then classified the direction of the results and strength and direction of the conclusions following predetermined criteria (Appendix A) [19−23] and established consensus.

#### 2.2.2. Training and pilot testing

At the pilot stage, two reviewers at each center (A.G., M.G., G.D., M.C., M.B., and B.P.) and three methods experts (L.H., R.M.F., and D.P.) familiarized themselves with published versions of each tool and their associated guidance documentation [7,8,16,24]. Next, reviewers independently piloted each tool on four SRs and then met to establish consensus. The teams undertook further pilot rounds and/or developed internal decision rules (e.g., for items with more subjective response options) as needed.

A pair of reviewers at each center then independently appraised each SR using the three tools in a random sequence to account for the expected increase in efficiency when applying the second and third tools in a series. Reviewers recorded the time taken to complete the appraisals, starting when they began reading the SR until all items had been completed. Reviewers applied all three tools for a given SR before moving on to the next. Although it is not recommended in practice [6,25], we tabulated an overall AMSTAR score as the sum of all 'yes' responses divided by the total number of items appraised for use in the utility analysis. After completing all appraisals, reviewers met to resolve discrepancies. The reviewers recorded the time taken to reach consensus, which started the moment the reviewers compared their appraisals and ended when a final decision was recorded.

### 2.3. Data analysis

We transferred all relevant data to SPSS Statistics (v. 24, International Business Machines Corporation, Armonk, NY), S-Plus (Version 8.2, TIBCO Software, Palo Alto, CA), or StatXact (v. 11, Cytel, Cambridge, MA) for analysis. We summarized the characteristics of the SRs descriptively.

We collapsed "yes" and "partial/probably yes" as well as "no" and "probably no" responses on the AMSTAR 2 and ROBIS tools before the analysis, but also performed analyses using all possible response categories. We calculated IRR and intercenter reliability for all items and overall ratings on each tool using Gwet's first-order agreement (AC1) statistics and 95% confidence intervals (CIs). For comparability to previous evaluations [9,10,12,14−16,26], we also calculated weighted Cohen's kappa statistics [27]. We interpreted an IRR $<0$ as poor, 0.0 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 to 0.99 as almost perfect [28]. We assessed the tools' utility (whether assessments could be used as an inclusion criterion for SRs in overviews without changing the results and/or conclusions) by testing for correlations between the results and conclusions for the primary outcome of each SR and reviewers' consensus appraisals on overall scores for each of the tools (significant at $P < 0.05$). To evaluate usability, we calculated the median (interquartile range [IQR]) time for reviewers to complete each tool and to reach consensus.

### 2.4. Differences from the protocol

We had planned a series of decision rules to determine the primary outcome of each SR [18], which were difficult to apply in practice; thus, we modified the hierarchy as described herein. We had not planned to collapse categories on the AMSTAR and ROBIS tools but did so for ease of comparison to previous evaluations [10,12,14].

## 3. Results

### 3.1. Characteristics of the systematic reviews

Appendix B shows the descriptive characteristics of the 30 SRs and the results and conclusions for their primary outcomes. In brief, the sample comprised 4 (13%) Cochrane SRs with meta-analysis and 2 (7%) with narrative synthesis, as well as 19 (63%) non-Cochrane SRs with meta-analysis and 5 (17%) with narrative synthesis. The SRs contained a median 8 RCTs (range 0 to 46) with 1,156 participants (range 0 to 37,655).

### 3.2. Inter-rater and intercenter reliability

#### 3.2.1. AMSTAR

Appendix C shows the IRR and intercenter reliability (Gwet's AC1) for each item on the three tools. Appendix D contains supplementary findings using all possible response options (rather than collapsed categories) for AMSTAR 2 and ROBIS, as well as Kappa values.

IRR was substantial to almost perfect across all centers for 8/11 (73%) items. The IRR ranged from moderate to almost perfect on item 3 (comprehensiveness of the search), moderate to substantial on item 4 (the status of publication used as an inclusion criterion) and slight to almost perfect

on item 8 (the quality of the included studies used appropriately in formulating conclusions).

Intercenter agreement was substantial to almost perfect for 6/11 (55%) items and ranged from moderate to substantial for items 2 (duplicate study selection and data extraction), 3 (comprehensiveness of the search), 4 (the status of the publication as an inclusion criterion), and 9 (appropriateness of the methods to combine findings). Agreement was moderate on item 8 (the quality of the studies used appropriately in formulating conclusions).

#### 3.2.2. AMSTAR 2

IRR was substantial to almost perfect across all centers for 9/16 items (56%), ranged from moderate to almost perfect on item 11 (appropriate statistical methods), and from fair to almost perfect on items 1 (included components of PICO), 4 (comprehensive search), 12 (impact of ROB), 13 (accounted for ROB in results), and 14 (explained heterogeneity). On item 7 (the list of excluded studies), IRR was slight at one center but almost perfect at the others. Agreement on the overall rating of confidence in the results ranged from slight to perfect.

Intercenter agreement was substantial to almost perfect across all between-center comparisons on 12/16 (75%) items. Agreement on item 6 (duplicate data extraction) was moderate for two between-center comparisons but almost perfect for the remaining comparison. Intercenter agreement was moderate across all comparisons for item 12 (the impact of ROB), ranged from slight to moderate (center A vs. B) for item 13 (account for ROB in results) and fair to moderate for item 14 (explained heterogeneity). Intercenter agreement on the overall rating of confidence in the results was moderate to substantial.

#### 3.2.3. ROBIS

Across all centers, IRR was substantial to almost perfect for 12/24 (50%) items and was substantial to almost perfect on all items at center C. Inter-rater agreement was fair to moderate for at least one center on signaling questions 1.3 (unambiguous eligibility criteria), 1.5 (restrictions based on information sources), 2.3 (search likely to retrieve eligible studies), 2.4 (search restrictions appropriate; both centers), 3.3 (all relevant results collected), 3.4 (ROB formally assessed), 4.4 (heterogeneity minimal or addressed), 4.6 (bias minimal or addressed), and overall ROB item A (interpretation addresses concerns). Inter-rater agreement was only slight for at least one center on items 4.2 (predefined analyses followed) and 4.5 (findings robust). Reviewer agreement was moderate to substantial or almost perfect on overall ratings for each domain except for domain 4. Agreement on the overall ROB in the review ranged from moderate to almost perfect.

Intercenter agreement was substantial to almost perfect on all between-center comparisons for 7/24 (62.5%) items. Across the remaining items, intercenter agreement was fair to moderate for two of three centers on items 1.5 (restrictions based on information sources), 2.2 (search beyond databases), 2.3 (search likely to retrieve eligible studies), 3.3 (all relevant results collected), 4.5 (findings robust), 4.6 (bias minimal or addressed), and overall ROB item C (avoid emphasizing results based on statistical significance). Intercenter agreement was slight for at least one between-center comparison on items 1.1 (predefined objectives), 2.1 (appropriate range of databases), 3.1 (data collection errors minimized), 4.4 (heterogeneity minimal or addressed). Intercenter agreement on overall domain ratings was fair for domain 1, slight to moderate for domain 2, fair to moderate for domain 3, and poor to moderate for domain 4. Agreement on the overall ROB in the review ranged from poor to moderate.

### 3.3. Usability

When completed first in a series, the median (IQR) time to complete the assessments was 15.7 (11.3) minutes for AMSTAR, 19.7 (12.1) minutes for AMSTAR 2, and 28.7 (17.5) minutes for ROBIS. The median (IQR) time to reach consensus was 2.6 (3.3) minutes for AMSTAR, 4.6 (5.3) minutes for AMSTAR 2, and 10.9 (10.8) minutes for ROBIS. Time data by center are in Appendix D.

### 3.4. Utility for reviewers

Table 1 shows the correlations of the overall appraisal for each tool and the results and conclusions of the SRs.

**Table 1.** Correlation between categorized results and conclusion for the primary outcome and the overall AMSTAR, AMSTAR 2, or ROBIS score

| | Spearman or Pearson correlation[a] (*P* value) | | | | | |
|---|---|---|---|---|---|---|
| | Center A | | Center B | | Center C | |
| Tool | Results | Conclusion | Results | Conclusion | Results | Conclusion |
| AMSTAR[b] | 0.006 (0.98) | 0.36 (0.05) | 0.09 (0.63) | 0.43 (0.02) | −0.10 (0.58) | 0.24 (0.20) |
| AMSTAR 2 | −0.04 (0.85) | −0.49 (0.006) | 0.09 (0.64) | −0.25 (0.19) | 0.01 (0.95) | −0.22 (0.25) |
| ROBIS | −0.35 (0.06) | −0.58 (0.0008) | −0.09 (0.64) | −0.19 (0.32) | 0.16 (0.39) | 0.04 (0.83) |

*Abbreviations*: AMSTAR, A MeaSurement Tool to Assess systematic Reviews; ROBIS, Risk Of Bias In Systematic reviews.

[a] We present the Spearman correlation for ordinal values (AMSTAR 2 and ROBIS) and the Pearson correlation for continuous values (AMSTAR).

[b] For the purposes of this study, we calculated an overall AMSTAR score by summing the number of 'yes' responses and dividing these by the total number of items in the tool (not including 'not applicable' items).

The results of the SRs were not significantly correlated with the overall AMSTAR, AMSTAR 2, or ROBIS appraisals. The overall AMSTAR 2 ($r = -0.49$, $P < 0.01$) and ROBIS ($r = -0.58$, $P < 0.01$) appraisals of center A showed moderate to strong inverse correlations with the conclusions of the SRs. The overall AMSTAR appraisal of center B showed a moderate correlation ($r = 0.43$, $P = 0.02$) with the conclusions of the SRs. These relationships were not observed with the appraisals of the other centers.

## 4. Discussion

In this study, pairs of reviewers with varying levels of experience across three centers applied AMSTAR, AMSTAR 2, and ROBIS to 30 SRs of healthcare interventions. Compared with AMSTAR, reviewers required slightly more time to complete AMSTAR 2, and substantially more for ROBIS. It took reviewers substantially longer to reach consensus for ROBIS than for AMSTAR and AMSTAR 2. IRR was highly variable for AMSTAR 2 and even more so for ROBIS. Intercenter agreement was lower and more variable for ROBIS than the two AMSTAR tools. Based on this variability, we could not draw clear conclusions about the utility of the tools for informing the inclusion of SRs in overviews because the correlations between the tools' rating and the results and conclusions of the SRs were not consistent across centers. For example, there was an inverse correlation between ROBIS appraisals and SR conclusions at center A but not at the two other review centers.

Similar to other studies [10,14,26], we found IRR to be highly variable across reviewer pairs, primarily for AMSTAR 2 and ROBIS. As others have previously observed [10,12,14,19], agreement seemed to be lowest and varied most substantially on items requiring high levels of subjective judgment. IRR may vary based on reviewers' familiarity with the tool [26], expertise in clinical appraisal [14,26], variability in how the guidance is interpreted and applied [26,29], and whether reviewers have worked together previously [10,29]. In our study, the reviewers at center C achieved substantial agreement for almost all items despite being relatively inexperienced in using the three tools. We found this difficult to explain (especially for ROBIS) but emphasize that a high degree of IRR does not infer 'correctness' [26]. It is conceivable that in an effort to achieve consensus, inexperienced reviewers might oversimplify items that are subjective and/or difficult to assess. Further evaluation is needed to better understand the effect of reviewer experience on the results of methodological quality or ROB appraisals.

The variable IRR that we observed was in spite of review teams performing at least one round of calibration exercises. In addition, our findings indicate that there may be marked variation in how the tools, especially AMSTAR and ROBIS, are interpreted and applied across centers. Indeed, AMSTAR has been criticized for unclear guidance on some items [30−32], which can lead to varying interpretations. ROBIS is accompanied by voluminous documentation [24], but its application requires considerable expertise and appears to rely heavily on subjective judgment [10,12,14]. The decision rules developed by each center revealed systematic differences in how the tools were applied. Owing to ambiguous guidance, these differences were most notable for ROBIS. For example, on item 2.1 (appropriateness of the search's databases/sources) one center required that authors search "at minimum MEDLINE, EMBASE, conference reports, and trial registers," whereas another applied less stringent criteria, requiring only "multiple databases and some unpublished literature." Explicit guidance on how to apply the tools to empty reviews was lacking. To assist readers in interpreting the findings of overviews, it will be important that authors provide a clear description of decision rules used in performing their assessments.

With respect to the individual tools, AMSTAR was the quickest to apply and pairs of reviewers obtained at least substantial agreement on most items. Our findings are similar to previous evaluations of AMSTAR [12,13,15,19,26], and we agree that the tool would be relatively easy to apply even for inexperienced reviewers [10,12]. However, intercenter reliability was consistently substantial for only about half (55%) of items. Reviewers appreciated the relatively unambiguous and concise available guidance for the newer AMSTAR 2. This may help to explain the substantial intercenter reliability that was achieved for most items. Reviewers still had difficulty with several items that required a greater degree of subjective judgment (e.g., items 11 to 14). In the case of one center, reviewers had difficulty even with some items that seem quite straightforward (e.g., item 7, the list of excluded studies, at center B). The reason for this is unclear and differs from the findings of Lorenz et al. [11], and the other two centers in our study that achieved substantial-to-almost perfect agreement on this item. Similarly, studies by Melchiors et al. [33] and Kang et al. [34] found that IRR was only fair for the corresponding AMSTAR item (item 5, the list of included and excluded studies provided). This may indicate some misinterpretation of the item, which further emphasizes the need for clear and concise guidance documentation for reviewers. Rating the overall confidence in the results of the review is a welcome addition in AMSTAR 2, but more guidance on how to ascertain this confidence may also be helpful because IRR and intercenter agreement were variable (potentially related to disagreement on other critical items within the tool). The available guidance was somewhat ambiguous about how to deal with the 'partial yes' responses in deciding whether SRs contained critical flaws. Nonetheless, the tool was relatively quick to apply, making it a practical option even for overviews with many SRs.

It is clear from the marked variation in IRR and relatively low intercenter reliability for more than half of the items that reviewers experienced difficulty in applying ROBIS. The relative ambiguity of available guidance documents and focus on subjective judgment meant that reviewers often felt that they were lacking in the necessary

expertise (especially content knowledge) to apply the tool. Similar to the findings of other evaluations [14], we found that the time spent fully understanding and deliberating over the 'probably' responses could be significant. The time to complete ROBIS is about double that of AMSTAR, which might reduce the tool's practicality. Because the application of ROBIS across review centers appears to be highly inconsistent, it is important that review teams consider completing their own assessments rather than 'borrowing' the previously completed assessments of others for their overviews. This will help to avoid misguided comparisons across reviews, which may be the result of disparate ratings of various author groups rather than actual differences in ROB across reviews. It seems unwise at this time to use ROB ratings obtained using ROBIS as an inclusion criterion for SRs in overviews because it is still unclear whether this might change the overview's findings.

### *4.1. Strengths and limitations*

To our knowledge, this is the only study to have concurrently tested and compared the reliability, usability, and utility of all three of AMSTAR, AMSTAR 2, and ROBIS. The generalizability of our evaluation is limited by the small sample of SRs of RCTs. Higher IRR on AMSTAR could have been related to reviewers' greater familiarity with this tool compared with the others. We evaluated the use of the three tools only for SRs of RCTs. It is not clear how our findings might have differed if SRs including nonrandomized studies were included in our sample. In addition, the type of expertise and background knowledge required by reviewers to appropriately use the tools is still not well understood. These uncertainties should be a focus of future research.

## 5. Conclusions

Compared with AMSTAR 2 and ROBIS, reviewers completed AMSTAR appraisals more quickly and with better agreement. Intercenter reliability was highest for AMSTAR 2, but ratings on the overall confidence in the results was variable. Both IRR and intercenter reliability were highly variable for ROBIS. Low levels of intercenter reliability, particularly on overall ratings of confidence or ROB, may limit readers' ability to interpret the ratings applied by various review groups. We could not draw any clear conclusions about potential relationships between reviewers' ratings on the tools and the results and conclusions of the SRs. Improved documentation and/or training resources are needed to facilitate consistent application of the tools.

## CRediT authorship contribution statement

**Michelle Gates:** Project administration, Formal analysis, Data curation, Writing - original draft, Writing - review & editing. **Allison Gates:** Conceptualization, Methodology, Project administration, Formal analysis, Data curation, Writing - original draft, Writing - review & editing. **Gonçalo Duarte:** Writing - review & editing. **Maria Cary:** Writing - review & editing. **Monika Becker:** Writing - review & editing. **Barbara Prediger:** Writing - review & editing. **Ben Vandermeer:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **Ricardo M. Fernandes:** Conceptualization, Methodology, Data curation, Writing - review & editing. **Dawid Pieper:** Conceptualization, Methodology, Data curation, Writing - review & editing. **Lisa Hartling:** Conceptualization, Methodology, Project administration, Writing - review & editing.

## Acknowledgments

The authors thank Jocelyn Shulhan-Kilroy for assistance with data extraction and verification.

Authors' contributions: A.G., B.V., D.P., R.M.F., and L.H. made substantial contributions to the conception and design of the study. Authors from the Canadian center (A.G., M.G., and L.H.) were responsible for project administration. A.G., M.G., G.D., M.C., B.P., and L.H. performed the risk of bias/quality appraisals. A.G. and M.G. collected other relevant data related to the studies. B.V. performed the statistical analyses. A.G., M.G., R.M.F., D.P., and L.H. discussed and interpreted the results. M.G. and A.G. wrote the first draft of the manuscript. All authors revised the manuscript critically for important intellectual content and approved the final version submitted.

## Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.04.026.

## References

[1] Hartling L, Chisholm A, Thomson D, Dryden DM. A descriptive analysis of overviews of reviews published between 2000 and 2011. PLoS One 2012;7:e49667.

[2] Pieper D, Buechter R, Jerinic P, Eikermann M. Overviews of reviews often have limited rigor: a systematic review. J Clin Epidemiol 2012; 65(12):1267−73.

[3] Li L, Tian J, Tian H, Sun R, Liu Y, Yang K. Quality and transparency of overviews of systematic reviews. Evid Based Med 2012;5(3): 166−73.

[4] Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco A, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. PLoS Med 2016;13(5):e1002028.

[5] Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.

[6] Pollock M, Fernandes RM, Becker L, Featherstone R, Hartling L. What guidance is available for researchers conducting overviews of

reviews of healthcare interventions? A scoping review and qualitative metasummary. Syst Rev 2016;5(1):190.
[7] Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008.
[8] Whiting P, Savović J, Higgins J, Caldwell D, Reeves B, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016;69:225−34.
[9] Pieper D, Buecheter RB, Li L, Prediger B, Eikermann M. Systematic review found AMSTAR, but not R (evised)-AMSTAR, to have good measurement properties. J Clin Epidemiol 2015;68:574−83.
[10] Pieper D, Puljak L, Gonzalz-Lorenzo M, Minozzi S. Minor differences were found between AMSTAR 2 and ROBIS in the assessment of systematic reviews including both randomized and nonrandomized studies. J Clin Epidemiol 2019;108:26−33.
[11] Lorenz RC, Matthias K, Pieper D, Wegewitz U, Morche J, Nocon M, et al. A psychometric study found AMSTAR 2 to be a valid and moderately reliable appraisal tool. J Clin Epidemiol 2019;114:133−40.
[12] Banzi R, Cinquini M, Gonzalez-Lorenzo M, Pecoraro V, Capobussi M, Miozzi S. Quality assessment versus risk of bias in systematic reviews: AMSTAR and ROBIS had similar reliability but differed in their construct and applicability. J Clin Epidemiol 2018;99:24−32.
[13] Gomez-Garcia F, Ruano J, Gay-Mimbrera J, Aguilar-Luque M, Sanz-Cabanillas JL, Alcalde-Mellado P, et al. Most systematic reviews of high methodological quality on psoriasis interventions are classified as high risk of bias using ROBIS tool. J Clin Epidemiol 2017;92:79−88.
[14] Bühn S, Mathes T, Prengel P, Wegewitz U, Ostermann T, Robens S, et al. The risk of bias in systematic reviews tool showed fair reliability and good construct validity. J Clin Epidemiol 2017;91:121−8.
[15] Perry R, Leach V, Davies P, Penfold C, Ness A, Churchill R. An overview of systematic reviews of complementary and alternative therapies for fibromyalgia using both AMSTAR and ROBIS as quality assessment tools. Syst Rev 2017;6(1):97.
[16] Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009;62:1013−20.
[17] Pollock M, Fernandes R, Becker L, Pieper D, Hartling L. Chapter V: overviews of reviews [draft]. In: Higgins J, Thomas J, Chandler J, Cumpston MS, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: Cochrane; 2020.
[18] Gates A, Gates M, Duarte G, Cary M, Becker M, Prediger B, et al. Evaluation of the reliability, usability, and applicability of AMSTAR, AMSTAR 2, and ROBIS: protocol for a descriptive analytic study. Syst Rev 2018;7(1):85.
[19] Pollock M, Fernandes RM, Hartling L. Evaluation of AMSTAR to assess the methodological quality of systematic reviews in overviews of reviews of healthcare interventions. BMC Med Res Methodol 2017;17:48.
[20] Tricco AC, Tetzlaff J, Pham B, Brehaut J, Moder D. Non-Cochrane vs. Cochrane reviews were twice as likely to have positive conclusion statements: cross-sectional study. J Clin Epidemiol 2009;62. 380-6e1.
[21] Lai NM, Teng CL, Lee ML. Interpreting systematic reviews: are we ready to make our own conclusions? A cross-sectional study. BMC Med 2011;9:30.
[22] Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. Health Technol Assess 2000;4:1−115.
[23] Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122−4.
[24] Whiting P, Savović J, Higgins J, Caldwell D, Reeves B, Shea B, et al. ROBIS guidance. 2017. Available at https://www.bristol.ac.uk/population-health-sciences/projects/robis/.
[25] Shea B, Dubé C, Moher D. Chapter 7. Assessing the quality of reports of systematic reviews: the QUORUM statement compared to other tools. In: Egger M, Smith GD, Altman DG, editors. Systematic Reviews in Health Care. London, UK: BMJ Books; 2008:122−39.
[26] Pieper D, Jacobs A, Weikert B, Fishta A, Wegewitz U. Inter-rater reliability of AMSTAR is dependent on the pair of reviewers. BMC Med Res Methodol 2017;17:98.
[27] Leibetrau A. Measures of association. Newbury Park, CA: Sage Publications; 1983.
[28] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.
[29] Hartling L, Hamm M, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments. AHRQ Methods for Effective Health Care. Rockville (MD): Agency for Healthcare Research and Quality; 2012.
[30] Burda BU, Holmer JK, Norris SL. Limitations of A Measurement tool to assess systematic reviews (AMSTAR) and suggestions for improvement. Syst Rev 2016;5:58.
[31] Faggion CM. Critical appraisal of AMSTAR: challenges, limitations, and potential solutions from the perspective of an assessor. BMC Med Res Methodol 2015;15:63.
[32] Wegewitz U, Weikert B, Fishta A, Jacobs A, Pieper D. Resuming the discussion of AMSTAR: what can (should) be made better? BMC Med Res Methodol 2016;16:111−33.
[33] Melchiors AC, Correr CJ, Venson R, Pontarolo R. An analysis of quality of systematic reviews on pharmacist health interventions. Int J Clin Pharm 2012;34(1):32−42.
[34] Kang D, Wu Y, Hu D, Hong Q, Wang J, Zhang X. Reliability and external validity of AMSTAR in assessing quality of TCM systematic reviews. Evid Based Complement Alternat Med 2012;2012:732195.