Journal of Clinical Epidemiology 114 (2019) 133−140

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

A psychometric study found AMSTAR 2 to be a valid and moderately reliable appraisal tool

Robert C. Lorenz[a,b,*], Katja Matthias[a], Dawid Pieper[c], Uta Wegewitz[d], Johannes Morche[a], Marc Nocon[a], Olesja Rissling[a], Jacqueline Schirm[a], Anja Jacobs[a]

[a]*Federal Joint Committee (Healthcare), Medical Consultancy Department, Gutenbergstr. 13, 10587 Berlin, Germany*
[b]*Division of Social and Preventive Medicine, University of Potsdam, Research Focus Cognitive Sciences, Am Neuen Palais 10, Potsdam 14469, Germany*
[c]*Witten/Herdecke University, School of Medicine, Faculty of Health, Evidence-based Health Services Research, IFOM − Institute for Research in Operative Medicine, Ostmerheimer Str. 200, 51109 Cologne, Germany*
[d]*Federal Institute for Occupational Safety and Health, Nöldnerstr.40-42, 10317 Berlin, Germany*

Accepted 24 May 2019; Published online 29 May 2019

Abstract

Objectives: The objectives of this study were to determine the interrater reliability (IRR) of assessment of multiple systematic reviews (AMSTAR) 2 for reviews of pharmacological or psychological interventions for the treatment of major depression, to compare it to that of AMSTAR and risk of bias in systematic reviews (ROBIS), and to assess the convergent validity between the appraisal tools.

Study Design and Setting: Two groups of four raters were each assigned one of two samples of 30 systematic reviews. All eight raters applied AMSTAR 2 to their sample. Each group also applied either AMSTAR or ROBIS. Fleiss' kappa and Gwet's AC_1 were calculated, and agreement between the tools was assessed.

Results: The median kappa values as a measure of IRR indicated a moderate agreement for AMSTAR 2 (median = 0.51), a substantial agreement for AMSTAR (median = 0.62), and a fair agreement for ROBIS (median = 0.27). Validity results showed a positive association for AMSTAR and AMSTAR 2 ($r = 0.91$) as well as ROBIS and AMSTAR 2 ($r = 0.84$). For the overall rating, AMSTAR 2 showed a high concordance with ROBIS and a lower concordance with AMSTAR.

Conclusion: The IRR of AMSTAR 2 was found to be slightly lower than the IRR of AMSTAR and higher than the IRR of ROBIS. Validity measurements indicate that AMSTAR 2 is closely related to both ROBIS and AMSTAR.

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

1. Introduction

Not only is the number of clinical trials exponentially increasing, the number of systematic reviews (SRs) has also been steadily growing [1]. They are furthermore of varying quality [2]. Decision makers need high-quality SRs to arrive at sound evidence-based health care decisions. A measurement tool for the assessment of multiple systematic reviews (AMSTAR) is a well-established tool [3,4] with good psychometric properties in terms of reliability and validity [5] for the methodological appraisal for SRs [6]. An updated version—AMSTAR 2—was developed and published in response to various criticisms (e.g., [7]) and constraints (e.g., only developed for SRs of RCTs) [8]. While both versions of AMSTAR focus on the appraisal of methodological quality, a recently published tool, risk of bias in systematic reviews (ROBIS), was specifically developed to assess of the risk of bias in SRs [9].

Because of the novelty of AMSTAR 2 and ROBIS, evidence for the psychometric properties of these tools remains scarce. Only a handful of studies have systematically investigated the interrater reliability (IRR) for AMSTAR 2 [8] and ROBIS [10-12]. Moreover, there is no published validation of AMSTAR 2. The present study therefore systematically investigates the IRR for AMSTAR 2 and compares it to AMSTAR and ROBIS. It furthermore examines the convergent validity between these instruments to show the relationship between the tools. As these instruments are most frequently applied in the context of guidelines or overviews of SRs, the reliability and validity are tested on SRs on a single research question.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest: The authors declare that they have no conflicts of interest.

* Corresponding author. Tel.: +49 30 275 838 352; fax: +49 30 275 838 305.

E-mail address: robert.lorenz@g-ba.de (R.C. Lorenz).

https://doi.org/10.1016/j.jclinepi.2019.05.028
0895-4356/

What is new?

Key findings

- Interrater reliability for assessment of multiple systematic reviews (AMSTAR) 2 was slightly lower in comparison to AMSTAR but higher in comparison to risk of bias in systematic reviews (ROBIS).
- AMSTAR 2 and ROBIS showed a high concordance with regard to the overall judgment.
- AMSTAR 2 scores are highly correlated with AMSTAR and ROBIS scores.

What this adds to what was known?

- Based on median kappa values, AMSTAR 2 interrater reliability has a moderate strength of agreement.
- AMSTAR 2 is a valid instrument for assessing the methodological quality of systematic reviews.

What is the implication and what should change now?

- We recommend an extended calibration phase before applying AMSTAR 2.

2. Material and methods

2.1. Study design

This cross-sectional study was conducted in the context of a project, for which a protocol has been registered on PROSPERO database (CRD42018110214). We selected 60 SRs on pharmacological or psychological interventions in the treatment of major depression in adults (Appendix A) and divided them into two samples. The first sample was independently assessed by a group of four raters using AMSTAR 2 and AMSTAR (AMSTAR group). The second was independently assessed by a different group of four raters using AMSTAR 2 and ROBIS (ROBIS group). The eight raters had various levels of experience. A pair of experienced reviewers and a pair of less experienced reviewer was identified in each group (Appendix B).

Before the appraisal began, each group completed a calibration phase in which the raters applied their respective appraisal tools to an SR that was not included in the samples. Each group discussed items upon which they disagreed separately and all eight raters discussed the AMSTAR 2 items of disagreement in the same manner.

Each of the two samples of 30 SRs was evenly divided into two blocs and the order in which the SRs in each bloc were to be rated was fixed. The raters assessed each bloc twice: first either with AMSTAR or ROBIS and then with AMSTAR 2. The remaining bloc was then appraised in reverse order (first AMSTAR 2, then AMSTAR or ROBIS). The type of intervention (pharmacological or psychological) was balanced across the samples and blocs. After completing the appraisal, each rater completed a questionnaire including—among others—questions regarding prestudy experience with appraisal tools (Appendix C).

2.2. Bibliographic search and sample selection

The systematic search (with a time window from 2012 to 2017) and subsequent study selection (Appendix A) identified 72 SRs. Of these results, 30 SRs were on psychological interventions and 42 on pharmacological. Thirty of the 42 pharmacological SRs were randomly selected and served together with the 30 psychological SRs as the sample for this methodological study (Appendix D).

2.3. Appraisal tools

AMSTAR 2 consists of 16 items plus the overall confidence rating in the review results. Items are evaluated either with "Yes" or "No" (items 1, 3, 5, 6, 10, 13, 14, and 16); with "Yes", "Partial Yes", or "No" (items 2, 4, 7, 8, and 9); or with "Yes", "No", or "No meta-analysis conducted" (items 11, 12, and 15). For the overall confidence rating, the response options "High", "Moderate", "Low," and "Critically low" were possible (Appendix E).

AMSTAR consists of 11 items, all of which are evaluated with "Yes", "No", "Can't answer", and "Not applicable". We used the AMSTAR version with the additional notes available on the AMSTAR web site (https://amstar.ca/docs/AMSTARguideline.pdf).

ROBIS was applied as intended except phase 1 was omitted (optional relevance assessment). ROBIS consists of 24 signaling questions (SQs) assigned to four domains plus a rating of the overall risk of bias in the review results. Responses to the SQ are as follows: "Yes", "Probably Yes", "Probably No", "No", and "No information". The risk of bias in each domain and in the overall rating is furthermore rated as "Low", "High", or "Unclear".

All ratings were collected in predefined Excel sheets with dropdown lists.

2.4. Data analysis

We calculated Fleiss' kappa (κ) and Gwet's AC_1 statistic to determine the IRR [13]. To this end, AMSTAR ratings were dichotomized into "Yes" and all other options. For AMSTAR 2, we calculated a linear weighted kappa for items with the response options "Yes", "Partial Yes", or "No" as well as for the overall confidence rating; for all other items, the kappa was not weighted. For ROBIS, we aggregated the SQ ratings into three categories: "Yes or Probably Yes", "Probably No or No", and "No information". Finally, we classified agreement as poor (<0), slight (0.01-0.2), fair (0.21-0.4), moderate (0.41-0.6), substantial (0.61-0.8), or almost perfect (0.81-1) according to Landis

and Koch's benchmark scale [14]. We further explored differences in IRR coefficients between the different reviewer pairs (using Cohen's kappa) for each instrument by calculating Wilcoxon signed-rank tests [15].

To measure convergent validity, reference judgments were generated by applying the majority rule similarly to previous studies (e.g., [10]). In case of an ambiguous result (e.g., two "Yes"-responses and two "No"-responses), the judgment of the most experienced rater in each group (appendix B) was taken as final judgment. These cases occurred for AMSTAR in 8.5% (28 of 330), for AMSTAR 2 in 10.4% (106 of 1020), and for ROBIS in 14% (122 of 870) of all ratings.

We applied two analytical approaches to these reference judgments. First, we assessed the agreement between the overall ratings using the overall confidence rating for AMSTAR 2 and the overall risk of bias rating for ROBIS. No overall rating exists for AMSTAR and the use of a summary score has been repeatedly discussed [6,16,17]. However, to achieve an approximation of an overall rating, we used a classification procedure based on summary score as previously applied [10]. Therefore, a scoring with 8-11 fulfilled items (i.e., "Yes" responses) was categorized as high quality, 4-7 fulfilled items as medium quality, and less than four fulfilled items as low quality. Agreement between AMSTAR classification and AMSTAR 2 overall confidence rating was assessed on this basis.

Second, we investigated the relationship between the numbers of items fulfilled by each SR with the three appraisal tools to achieve a more differentiated picture (although this score is arbitrary). To this end, we used the aforementioned "yes" score for AMSTAR. For AMSTAR 2, we counted each "Yes" response with 1 scoring unit and each "Partial Yes" with 0.5 scoring units. For ROBIS, number of "Yes or Probably Yes"-responses to the 24 SQ was determined. Finally, ordinal Spearman rank correlation coefficients were calculated for the "yes" scores.

Data were analyzed and graphically prepared using statistic software packages of SPSS (IBM Corporation, Chicago, IL) and R (R Core Team). The latter was specifically used for calculating agreement coefficients [13,18].

3. Results

Forty-two of the 60 SRs included only RCTs and 18 SRs included both RCTs and non-RCTs. Four SRs were Cochrane Reviews. The results of κ statistics are reported for IRR. Results of Gwet's AC_1 statistics are primarily reported in the appendices.

3.1. Interrater reliability

For AMSTAR 2, we calculated the IRR based on both groups (60 SRs) resulting in a κ range across the items from 0.15 to 0.8 and a median of 0.51. This median indicates a moderate agreement (Table 1, Figure 1, and Appendix F). The κ of the overall confidence rating was 0.42 (95% CI: 0.25-0.59). A calculation of the IRR for AMSTAR 2 by group revealed no significant difference between the groups (Appendix G).

For AMSTAR (30 SR), the κ ranged between 0.16 and 0.88 with a median of 0.62. The median value indicates a substantial agreement (Figure 1 and Appendix F).

The ROBIS (30 SR) kappa values ranged from 0.03 to 0.85 for the SQ and from 0.23 to 0.38 for the domain judgments. The κ of the overall risk of bias was 0.27 (95% CI: −0.01 to 0.54). The median across all ROBIS SQ and domains was 0.27 indicating a fair strength of agreement (Figure 1 and Appendix F).

We further explored the impact of prestudy experience in the appraisal of SRs and found that the IRR did not systematically vary between the experienced reviewer pairs in each group for any of the appraisal tools (all $P > 0.05$, Appendix H).

3.2. Convergent validity

Considering the reference judgments for the overall ratings, the included SRs were of rather low quality. Analysis with AMSTAR 2 categorized the quality of 53 SRs out of 60 as critically low, 4 as low, 1 as moderate, and 2 as high. Based on AMSTAR classification, 6 SRs out of 30 were categorized as low quality, 15 as medium, and 9 as high. For ROBIS, the risk of bias of 26 SRs out of 30 was judged as high, 3 as low, and 1 as unclear. All SRs that were not determined to be of critically low quality by AMSTAR 2 were classified as high quality by AMSTAR, indicating some degree of concordance. Agreement between AMSTAR 2 and ROBIS was high, and almost all SRs ($n = 26$) were judged as being of critically low quality and at high risk of bias (Figure 2).

As the results of overall rating validation showed a floor effect (i.e., vast majority of SRs were rated with the lowest judgment level according to AMSTAR 2 and ROBIS overall ratings), the aggregated "yes" scores for each instrument may provide better differentiation. The Spearman rank correlation coefficient of the "yes" scores between AMSTAR and AMSTAR 2 indicated a strong positive association ($r = 0.91$, $P < 0.001$). Moreover, the correlation between AMSTAR 2 and ROBIS "yes" scores was also positively strong ($r = 0.837$, $P < 0.001$).

4. Discussion

4.1. Main findings

In this study, we investigated reliability and validity of the new appraisal tool AMSTAR 2 and compared it to AMSTAR and ROBIS. Based on our sample of SRs of pharmacological or psychological interventions for the treatment of major depression, the IRR of AMSTAR 2 was slightly lower than that of AMSTAR but higher compared to

Table 1. Overview of agreement and interrater reliability (Fleiss' kappa and Gwet's AC_1) for each AMSTAR 2 item

Item	pa	κ [95% CI]	AC_1 [95% CI]
1. Components of PICO question?	0.83	0.15 [0.03; 0.27]	0.79 [0.68; 0.89]
2. Review protocol?	0.92	0.74 [0.61; 0.88]	0.89 [0.81; 0.97]
3. Explanation of study design?	0.84	0.36 [0.15; 0.58]	0.79 [0.68; 0.9]
4. Comprehensive literature search strategy	0.78	0.39 [0.24; 0.54]	0.56 [0.47; 0.65]
5. Study selection in duplicate?	0.86	0.71 [0.58; 0.84]	0.73 [0.6; 0.86]
6. Data extraction in duplicate?	0.9	0.8 [0.69; 0.91]	0.8 [0.69; 0.91]
7. List of excluded studies and justify the exclusions?	0.87	0.64 [0.49; 0.79]	0.81 [0.71; 0.91]
8. Study characteristics	0.79	0.37 [0.23; 0.5]	0.6 [0.48; 0.71]
9. Satisfactory technique for assessing risk of bias?	0.79	0.53 [0.38; 0.68]	0.6 [0.47; 0.74]
10. Sources of funding for each study?	0.8	0.51 [0.35; 0.66]	0.66 [0.51; 0.81]
11. Appropriate methods?	0.72	0.57 [0.44; 0.69]	0.59 [0.48; 0.7]
12. Assess potential impact of risk of bias on the results?	0.78	0.66 [0.54; 0.77]	0.67 [0.57; 0.78]
13. Account for risk of bias when interpreting/discussing?	0.76	0.51 [0.36; 0.66]	0.54 [0.39; 0.68]
14. Satisfactory explanation for and discussion of any heterogeneity?	0.66	0.31 [0.16; 0.46]	0.32 [0.17; 0.48]
15. Publication bias (small sample bias) assessed and discussed?	0.84	0.75 [0.65; 0.86]	0.77 [0.68; 0.87]
16. Potential sources of conflict of interest?	0.8	0.33 [0.1; 0.57]	0.71 [0.59; 0.83]
Overall confidence rating	0.92	0.42 [0.25; 0.59]	0.89 [0.83; 0.95]

Abbreviations: AMSTAR, assessment of multiple systematic reviews; AC_1, Gwet's AC_1 statistic; CI, confidence interval; κ, Kappa coefficient (Fleiss' kappa); pa, percent agreement.

ROBIS. The median kappa value across all items of AMSTAR 2 indicated moderate agreement. Specifically, the agreement was found to be substantial for 6 items, moderate for 4 items, fair for 5 items, and slight for one item. Overall confidence rating showed a moderate agreement.

For AMSTAR, the median kappa value indicated a substantial agreement (3 almost perfect, 4 substantial, 3 moderate, and one item slight agreement). For ROBIS, the median kappa value indicated fair agreement (one almost perfect, 2 substantial, 5 moderate, 13 fair, and 8 items slight agreement). The overall risk of bias showed a fair agreement.

Results of the convergent validity showed a positive association for AMSTAR and AMSTAR 2 as well as ROBIS and AMSTAR 2. AMSTAR 2 and ROBIS showed a high concordance for the overall rating.

4.2. Our findings in context

The original publication of AMSTAR 2 showed a moderate to substantial IRR for most items [8]. Here, we observed a moderate to substantial agreement in 10 items and fair agreement in 5 items. The low IRR of item 1 (PICO question) stands out. The kappa value of 0.15 indicated a poor agreement; however, Gwet's AC_1 showed a substantial agreement (0.79). This difference is driven by a skewed distribution (i.e., 89% ''Yes'' responses and 11% ''No'' responses). Kappa punishes disagreements between raters in skewed distributions heavily, while Gwet's AC_1 is more stable in these cases [13,18,19]. Given that the percent agreement is 83%, the kappa value here may be biased. A comparable finding was also present for the items 3 (explanation of study design), 8 (PICO details), 16 (conflict of interest of SR) and, even more noteworthy, the overall confidence rating (kappa: 0.42, Gwet's AC_1: 0.89). Thus, the IRR may be underestimated. It should be noted here that Gwet's AC_1 statistics always lead to higher values than kappa statistics.

Our result of substantial agreement for most of the AMSTAR items is in line with the results of a previous reliability review for AMSTAR [5]. The IRR for item 3 (comprehensive literature search) and item 11 (conflict of interests) was lower in the current paper.

Our IRR results for ROBIS support the findings of fair agreement by Bühn et al. (16 SRs) [11] and an overview of SRs (15 SRs) for the therapy of fibromyalgia that applied ROBIS as appraisal tool by Perry et al. [20]. By contrast, Banzi et al. (31 SRs) [10], Gomez-Garcia et al. (139 SRs) [12], and Tao et al. (19 SRs) [21] found a higher IRR. Though the evidence of IRR for ROBIS is heterogeneous, it still seems to be lower in comparison to AMSTAR.

The possibility that an individual rater's degree of influence may influence the IRR is a matter of discussion. A previous study [11] showed that more experience is associated with lower (AMSTAR and ROBIS SQ) and higher IRR (ROBIS domain rating). In the present study, no influence of experience on IRR was observed. Thus, a systematic investigation of the impact of experience on IRR is needed.

We assessed convergent validity by investigating the concordance of the overall rating of AMSTAR 2 and ROBIS as well as a classification of the summary score of AMSTAR. Both AMSTAR 2 and ROBIS classified most of the

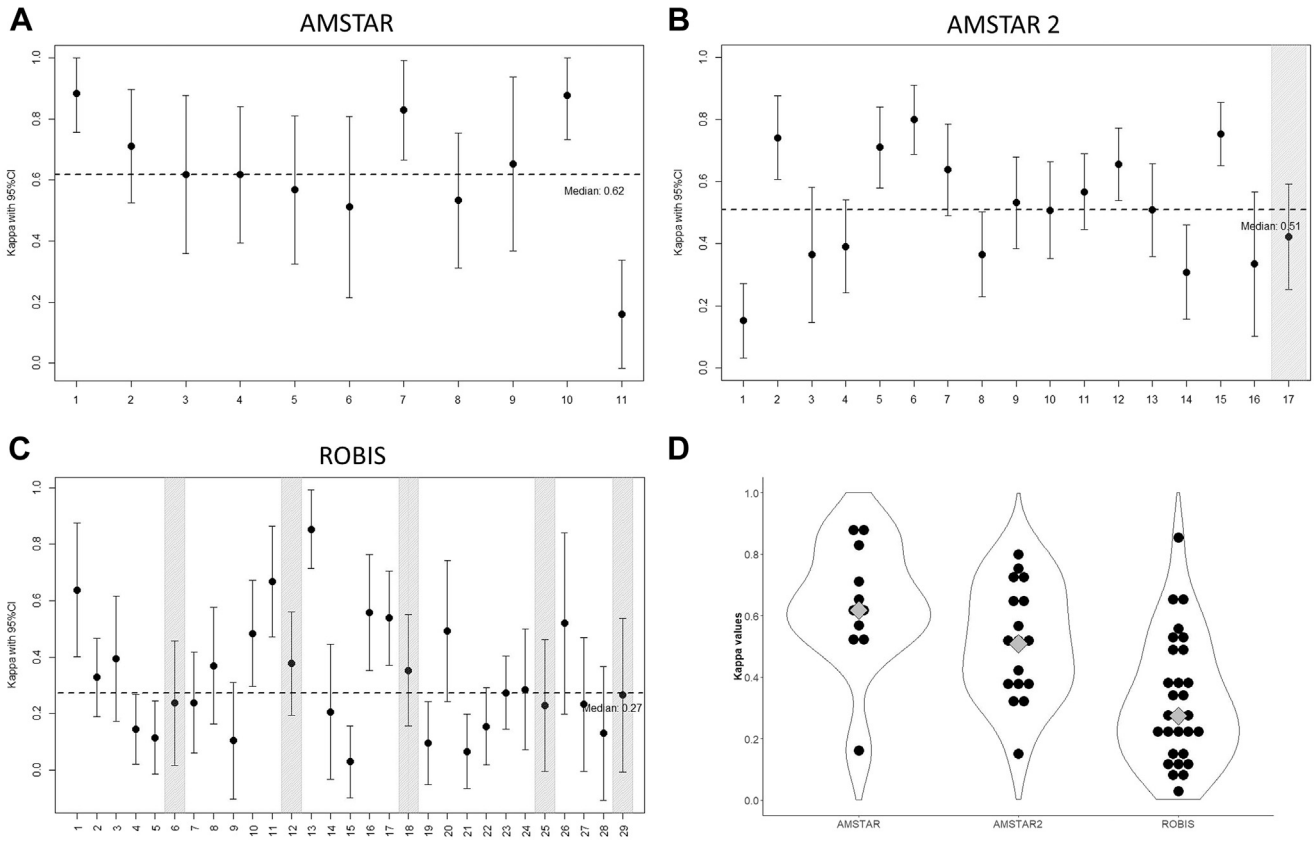

Fig. 1. Results of the interrater reliability of the three appraisal tools. The three investigated appraisal tools were AMSTAR (A), AMSTAR 2 (B), and ROBIS (C), and their kappa statistics as a measure of interrater reliability for each item are shown. (A-C) The items of the corresponding appraisal tools are shown on the x-axis of each panel. The kappa values with a 95% confidence interval (error bars) are depicted on the y-axis. Solid black circles represent Fleiss' kappa values for the four raters. Gray-shaded areas represent overall ratings (for ROBIS and also for each domain). The confidence intervals for AMSTAR 2 are narrower because kappa values are based on 60 SRs, whereas AMSTAR and ROBIS ratings were based on 30 SRs (see study design for details). (D) Violin plots of the kappa values of the three appraisal tools. Solid black circles represent kappa values of each item. Gray diamonds indicate median values of the corresponding appraisal tools. AMSTAR, assessment of multiple systematic reviews; ROBIS, risk of bias in systematic reviews; SRs, systematic reviews.

included SRs with the lowest quality category (floor effect), leading to a high agreement. According to AMSTAR 2, critically low quality is reached when two or more critical items are rated with "no" (e.g., review protocol and a list of all excluded studies in text form) [8]. For ROBIS, some concerns in one of the four domains without any discussion of these concerns may lead to devaluation to a high risk of bias overall [9]. Thus, both tools have a relatively conservative classification. By contrast, when applying the previously used classification to the AMSTAR summary score [10], even three "no" responses are acceptable for the category high quality. Therefore, the low concordance between AMSTAR and AMSTAR 2 is probably related to the lower requirements for a high-quality judgment when applying the AMSTAR classification.

As a second assessment for the convergent validity, we investigated the relationship of the summary scores between the appraisal tools and found a strong association between AMSTAR and AMSTAR 2 as well as between AMSTAR 2 and ROBIS. These findings are in line with a previous study [11]. AMSTAR and AMSTAR 2 are focused on methodological quality, whereas ROBIS is tailored to assess the risk of bias (evaluation of the internal validity). The two constructs assess different aspects: reviews with high methodological quality may still be judged as having high a risk of bias [10]. By contrast, SRs with methodological or reporting weaknesses may be rated as having a low risk of bias by ROBIS if the authors address all concerns in the review conclusions. In our sample, low methodological quality was associated with a high risk of bias, which may be driven by the low methodological quality of the investigated SRs (methodological quality may only partially explain the risk of bias). Thus, more research with SRs of high methodological quality is needed.

4.3. Strengths and limitations

To the best of our knowledge, this is the first study to apply AMSTAR 2 under conditions that are comparable to the intended application: all SRs were related to a single PICO question and the results can be used for the creation of an overview of SRs. Furthermore, the present study

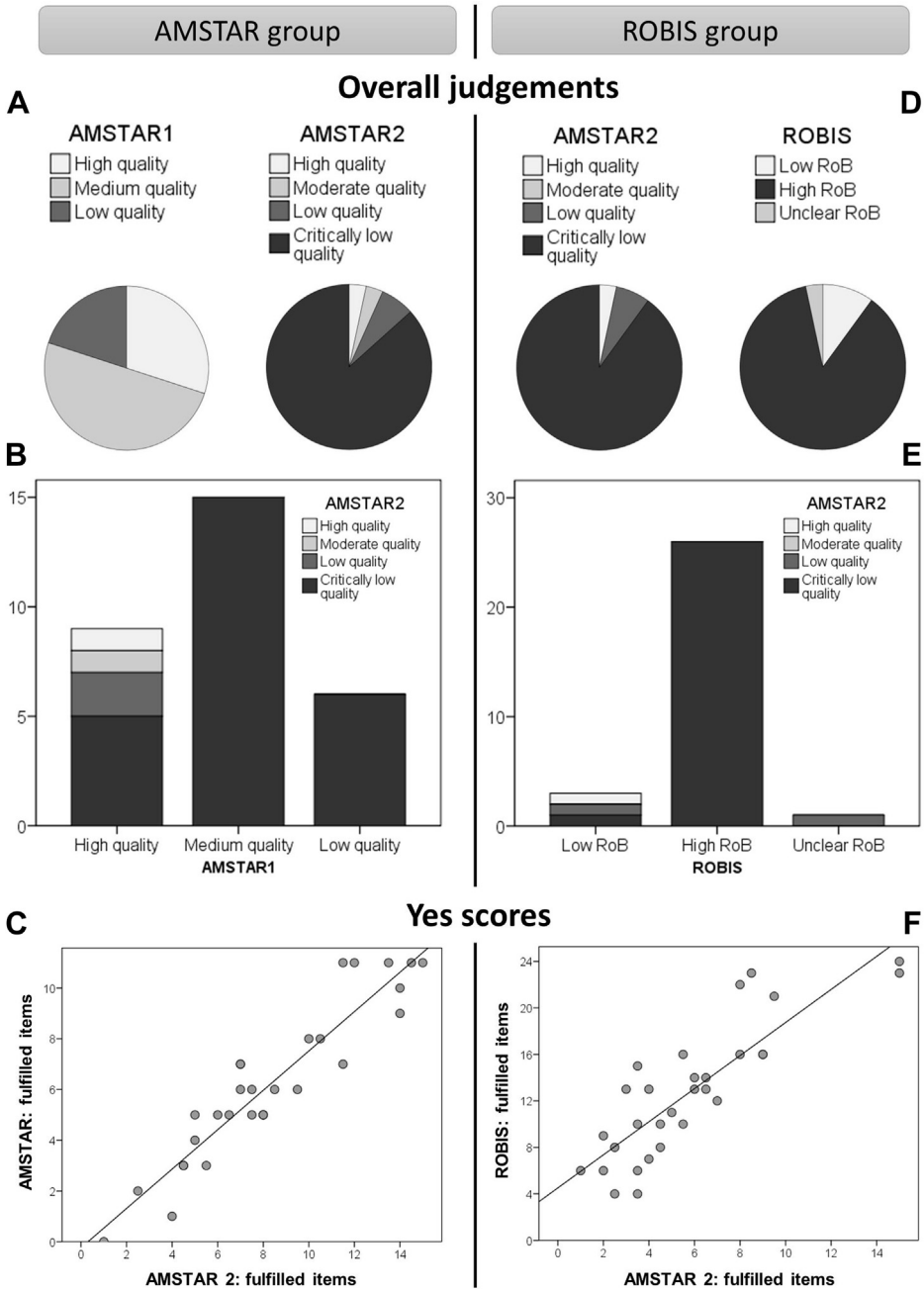

Fig. 2. Validity results for the two groups (AMSTAR and ROBIS). Panels A and D show pie charts of the overall judgments of the two SR samples (30 SRs per sample). Panels B and E show the stacked bar charts for the same results. To better differentiate between the SR with the lowest quality judgment, ''Yes'' scores (aggregated fulfilled items, i.e., answered ''Yes'') were generated and correlated for each appraisal tool as shown in panels C and F. AMSTAR, assessment of multiple systematic reviews; ROBIS, risk of bias in systematic reviews; SRs, systematic reviews.

comprised a relatively large sample (60 SRs) that was appraised by four raters.

The study has also some limitations. First, we were not able to apply AMSTAR and ROBIS to all SRs. Owing to time constraints, each rater applied two appraisal tools. Previous studies have directly compared AMSTAR and ROBIS [10,11].

Second, the impact of experience on IRR may be multifactorial. Here, we operationalized the experience by the number of reviews assessed with any appraisal tool, which is mainly related to experience with AMSTAR and not AMSTAR 2 or ROBIS (almost all reviewers had no experience with these tools). Another factor may be the working experience at the same institution; six of the eight reviewers worked at the same institution, not all reviewers were completely independent. A previous study has shown a greater agreement of reviewers within one study center than between multiple centers [22].

Third, the experience on AMSTAR may also have an impact on the novelty of the updated AMSTAR 2. Because

the experience of the reviewers was mainly related to AMSTAR (and therefore somehow also to AMSTAR 2), ROBIS may be the most unfamiliar instrument of these three, which may have an impact on the IRR.

Fourth, the calibration phase was completed on the basis of only one SR and may be too short to harmonize the assessment between the raters.

Fifth, the use of summary scores from appraisal tools has been a matter of much discussion in the literature [16,17,23,24]. We used these scores solely descriptively to show convergent validity, which is justifiable.

Sixth, to reach consensus between the ratings of the reviewers, we applied the majority rule. In cases where results were ambiguous (about 10%), we used the rating of the most experienced reviewer, which may slightly bias the validity results.

5. Conclusion

Given the increasing number of SRs and the high number of SRs of suboptimal quality [2], quality appraisal tools for SRs are of critical importance. The psychometric properties of appraisal tools are an important consideration in choosing an adequate instrument.

AMSTAR 2 is the updated version of AMSTAR and the application of AMSTAR may soon be outdated. In terms of IRR, AMSTAR 2 kappa values do not reach the high level of AMSTAR. ROBIS, on the other hand, focuses more on the risk of bias than on methodological quality and the assessment is much more detailed (29 items compared to 17 items for AMSTAR 2). Lower IRR values for ROBIS suggest that ratings are more variable. The validity results indicate that AMSTAR 2 is closely related to both ROBIS and AMSTAR and therefore both are valid instruments. We recommend an extended calibration phase before applying the tools.

CRediT authorship contribution statement

Robert C. Lorenz: Conceptualization, Methodology, Formal analysis, Project administration, Writing - original draft, Writing - review & editing. **Katja Matthias:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Dawid Pieper:** Conceptualization, Methodology, Writing - review & editing. **Uta Wegewitz:** Writing - review & editing. **Johannes Morche:** Writing - review & editing. **Marc Nocon:** Writing - review & editing. **Olesja Rissling:** Writing - review & editing. **Jacqueline Schirm:** Writing - review & editing. **Anja Jacobs:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing.

Acknowledgments

The authors thank Lydia Jones for assistance with language editing and proofreading.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.05.028.

References

[1] Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? Plos Med 2010; 7(9):e1000326.
[2] Ioannidis JP. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q 2016; 94(3):485−514.
[3] Pieper D, Buechter R, Jerinic P, Eikermann M. Overviews of reviews often have limited rigor: a systematic review. J Clin Epidemiol 2012; 65:1267−73.
[4] Pollock M, Fernandes RM, Becker LA, 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.
[5] Pieper D, Buechter 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.
[6] 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.
[7] 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.
[8] 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.
[9] Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, 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.
[10] Banzi R, Cinquini M, Gonzalez-Lorenzo M, Pecoraro V, Capobussi M, Minozzi 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.
[11] Buhn 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.
[12] 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.
[13] Gwet KL. Handbook of inter-rater reliability: the definitve guide to measuring the extent of agreement among raters. 4th ed. Gaithersburg: Advanced Analytics, LLC; 2014.
[14] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.
[15] Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull 1945;1(6):80−3.
[16] Burda BU, Holmer HK, Norris SL. Limitations of a Measurement Tool to Assess Systematic Reviews (AMSTAR) and suggestions for improvement. Syst Rev 2016;5:58.
[17] 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.
[18] Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61(Pt 1): 29−48.

[19] Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 2013; 13:61.

[20] 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.

[21] Tao H, Zhang Y, Li Q, Chen J. Methodological quality evaluation of systematic reviews or meta-analyses on ERCC1 in non-small cell lung cancer: a systematic review. J Cancer Res Clin Oncol 2017; 143(11):2245−56.

[22] Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol 2013;66:973−81.

[23] Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282:1054−60.

[24] Pieper D, Koensgen N, Breuing J, Ge L, Wegewitz U. How is AMSTAR applied by authors - a call for better reporting. BMC Med Res Methodol 2018;18:56.