ELSEVIER

Journal of Clinical Epidemiology 108 (2019) 26−33

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Minor differences were found between AMSTAR 2 and ROBIS in the assessment of systematic reviews including both randomized and nonrandomized studies

Dawid Pieper[a,*], Livia Puljak[b], Marien González-Lorenzo[c,d], Silvia Minozzi[e]

[a]*Institute for Research in Operative Medicine, Evidence-based Health Services Research, Faculty of Health, School of Medicine, Witten/Herdecke University, Ostmerheimer Str, Cologne 200 51109, Germany*

[b]*Catholic University of Croatia, Ilica 242, Zagreb 10000, Croatia*

[c]*Department of Biomedical Sciences, Humanitas University, Milan, Italy*

[d]*IBD Center, Humanitas Clinical and Research Center, Milan, Italy*

[e]*Department of Biomedical Sciences for Health, University of Milan, Milan, Italy*

Accepted 5 December 2018; Published online 10 December 2018

Abstract

Objective: To compare A Measurement Tool to Assess Systematic Reviews (AMSTAR 2) with a tool to assess risk of bias in systematic reviews (ROBIS) in terms of validity, reliability, and applicability.

Study Design and Setting: We analyzed 30 systematic reviews (SRs) that included randomized and nonrandomized studies, with Cochrane and non-Cochrane SRs sampled in 1:1 ratio. Four reviewers assessed independently all 30 SRs with AMSTAR 2, followed by ROBIS. We calculated Fleiss' Kappa as a measure of inter-rater reliability (IRR) across 4 raters.

Results: The IRR for scoring the overall confidence in the SRs with AMSTAR 2 and the overall domain in ROBIS was fair (AMSTAR 2: $\kappa = 0.30$, 95% [confidence interval] CI: 0.17 to 0.43; ROBIS: $\kappa = 0.28$, 95% CI: 0.13 to 0.42). AMSTAR 2 confidence in review ratings strongly correlated with the overall domain rating in ROBIS (Spearman $r_s = 0.84$). Mean time for scoring AMSTAR 2 was slightly higher than for ROBIS (18 vs. 16 min), with huge differences between the reviewers.

Conclusion: Both AMSTAR 2 and ROBIS can be applied to SRs including both randomized controlled trials (RCTs) and non-RCTs. Measurement properties of ROBIS seemed not to be much different when comparing with other studies that include only SRs of RCTs.

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

1. Introduction

Systematic reviews (SRs) can provide the highest level of evidence and inform evidence-based decision-making in health care. A necessary prerequisite is that they are based on a sound methodology to avoid bias. Therefore, it is important that SRs undergo critical appraisal. There are several tools that could be used for that. In the last decade, A Measurement Tool to Assess Systematic Reviews (AMSTAR) [1] has become the most common tool to assess the methodological quality of SRs [2,3]. It was claimed that AMSTAR has good measurement properties in terms of validity, reliability, and applicability [4]. Nevertheless, AMSTAR has also been subject to criticism by different authors [5−7].

Partly as a result of this, AMSTAR has been updated in 2017. The new tool, called AMSTAR 2, has been designed to enable a more detailed assessment of SRs [8]. Prior research has indicated that including nonrandomized controlled trials (non-RCTs) into SRs of therapeutic interventions is challenging [9]. AMSTAR 2 has also been developed to assess the methodological quality of SRs that include randomized controlled trial (RCT) or non-RCTs of health care interventions, or both. This is an extension over

Conflicts of interest: The authors declare to have no conflicts of interest.

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

Authors' Contributions: All authors were involved in designing the study, acquisition of data, and interpretation of data. DP performed all analyses and drafted the article. All authors revised it critically for important intellectual content and approved the finally submitted version.

* Corresponding author. Tel.: +49 221 98957-40; fax: +49 221 98957-30.
E-mail address: Dawid.pieper@uni-wh.de (D. Pieper).

https://doi.org/10.1016/j.jclinepi.2018.12.004
0895-4356/© 2018 Elsevier Inc. All rights reserved.

What is new?

Key findings

- A Measurement Tool to Assess Systematic Reviews (AMSTAR 2) confidence in review ratings strongly correlated with the overall domain rating in risk of bias in systematic reviews (ROBIS), whereas the interrater reliability for the overall judgment was fair for both, AMSTAR 2 and ROBIS.
- Both, AMSTAR 2 and ROBIS, are regarded as an improvement compared to AMSTAR with respect to assessment of methodological quality/risk of bias in systematic reviews.
- Interrater reliability was higher in reviewers who have worked in prior projects together.

What this adds to what is known?

- This is the first validation study for AMSTAR 2.
- Both, AMSTAR 2 and ROBIS, can also be applied to judge the methodological quality/risk of bias of systematic reviews including both, randomized and nonrandomized controlled trials.

What is the implication, what should change now?

- Authors are advised to perform a calibration exercise before using one of these tools.

AMSTAR that has originally only been designed to evaluate SRs of RCTs of health care interventions [1].

In 2016, another tool to assess the risk of bias in SRs called risk of bias in systematic reviews (ROBIS) has been published [10]. The ROBIS tool focuses on ROBIS, while the AMSTAR 2 tool focuses on methodological quality. Although both concepts are strongly related to each other, they cannot be treated equally [11]. Moreover, ROBIS can be applied to all forms of SRs [10].

So far only two studies comparing ROBIS with the older AMSTAR have been published [12,13]. To our knowledge there is no published study comparing the ROBIS with the newly developed AMSTAR 2 in the context of SRs including non-RCTs. This study aims to report on a first experience with AMSTAR 2 and compare it with ROBIS when assessing SRs that include both RCTs and non-RCTs while assessing validity, reliability and applicability.

2. Methods

This is a cross-sectional study. There was an unpublished a priori protocol for this study (see Appendix A).

2.1. Sample selection

To meet eligibility criteria, all SRs needed to include both RCTs and non-RCTs. Furthermore, they needed to investigate an intervention. Only SRs in English were considered. In total, we aimed to include 30 SRs, with Cochrane reviews (CRs) and non-Cochrane reviews (nCRs) sampled on a 1:1 ratio (i.e., each group included 15 reviews). This was done as prior research indicated that CRs have a higher methodological quality and reporting quality than nCRs [14−16].

We used two cohorts of SRs from our former projects where we knew that a sufficient number of SRs including also non-RCTs would be eligible. The first project dealt with SRs in the field of anesthesiology and pain, and included only nCRs [17]. The second project investigated characteristics of controlled before-after and interrupted time series studies (i.e., non-RCTs) included in CRs [18]. From each sample 15 SRs were selected randomly using the RAND function in Excel.

2.2. Assessment tools

Both tools were applied as intended by the developers. ROBIS consists of four domains: (1) study eligibility criteria, (2) identification and selection of studies, (3) data collection and study appraisal, and (4) synthesis and findings and a final overall judgment domain. Ratings of low, high, or unclear ROBIS are applied to each domain and overall.

AMSTAR 2 consists of 16 items. Items can be either answered with yes or no (items 1, 3, 5, 6, and 10-16), or yes, partial yes, or no (items 2, 4, 7, 8 and 9), while the reviewers may also choose an answer of not applicable for items 11, 12, and 15. An overall rating of high, moderate, low, or critically low quality is given to judge the overall confidence in the results of the review.

2.3. Reviewers

Four reviewers assessed independently all 30 SRs. All of them have a high level of experience in conducting SRs. Furthermore, all of them took part in former studies evaluating the methodological quality of SRs. All but one reviewer have already applied ROBIS before, while all have used AMSTAR 2 for their first time. No reviewer was formally trained in one of the tools. All SRs were assessed in the same order by all four reviewers. The order was randomized on a 1:1 ratio (i.e., CR, nCR, CR, …). For each included SR, assessments started with AMSTAR 2 followed by ROBIS. Three reviewers were naïve to the content of all SRs, while one reviewer (DP) has included some of the SRs while screening against eligibility criteria more than 2 years ago.

2.4. Data analysis

All data were collected in Excel sheets. To measure construct validity we converted the confidence in the SR rating in AMSTAR 2 and the overall domain in ROBIS into numerical values. For AMSTAR 2, values ranged from 1 (critically low quality) to 4 (high quality), while values for ROBIS ranged from 1 (high ROBIS) to 3 (low risk of bias). As no consensus procedure took place between the reviewers, we calculated a mean score across four raters for each review for both assessment tools. The obtained scores were used to calculate Spearman's rank correlation coefficient r_S.

We calculated Fleiss' Kappa (κ) as a measure of interrater reliability (IRR) for each AMSTAR 2 question and for each ROBIS domain and signaling question between all four reviewers [19]. In addition, we calculated Gwet's AC_1 statistic [20]. We classified agreement as poor ($\leq$0.00), slight (0.01-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), almost perfect (0.81-1.00) relying on accepted approaches [21].

We matched similar domains, where possible, assessed by both AMSTAR 2 and ROBIS to explore the concurrent validity of the two tools. We defined ten domains as fully overlapping, as we considered that the same content was addressed by one question of AMSTAR 2 and one or more questions of ROBIS. We defined two domains as partially overlapping because the two tools addressed the same content (appropriateness of restriction of eligibility criteria and robustness of the results) but defined it slight differently. We have dichotomized ratings for AMSTAR 2 and ROBIS (yes vs. no). As no consensus procedure took place, we considered the judgments of most of the raters as the reference judgment. We used Gwet's AC_1 statistic as an agreement coefficient.

To investigate applicability, we recorded the time to complete each tool for each SR using a digital chronograph.

In subgroup analyses, we investigated differences between CRs and nCRs. Furthermore, we repeated our analyses for pairs of reviewers against the background that this might have an influence on the results [22], and as two reviewers already took part together in a similar study [13].

3. Results

3.1. Review characteristics

We included 30 SRs (15 CRs and 15 nCRs) published between 2012 and 2016 (see Appendix B). The number of included studies in the SRs ranged from 2 to 118. More than half (18/30) of the SRs performed meta-analysis. Characteristics of the included SRs can be found in Appendix C.

3.2. Validity

AMSTAR 2 confidence in review ratings strongly correlated with the overall domain ratings in ROBIS (Fig. 1). The Spearman's rank correlation coefficient r_s between both tools was 0.84 (P = 0.00). This correlation was higher than for each of the single reviewer (r_s range 0.63 to 0.76).

3.3. Interrater reliability

For AMSTAR 2, perfect IRR was observed for item 2 (κ = 1.00, 95% [confidence interval] CI: 1 to 1). For items 3, 6, 10, 11a, and 12, IRR was substantial (Table 1). Lowest IRR was obtained for item 8 (κ = 0.09, 95% CI: 0.06 to 0.23), indicating only a slight agreement. The IRR for scoring the overall confidence in the SR was fair (κ = 0.30, 95% CI: 0.17 to 0.43).

For ROBIS, only domain 1 was found to have a moderate IRR with κ = 0.42 (95% CI: 0.27 to 0.57). All other domains only demonstrated a fair IRR (Table 2). The IRR for the overall domain was fair as indicated by κ = 0.28 (95% CI: 0.13 to 0.42). Appendix D presents IRR for signaling questions.

3.4. Matching analysis

Most fully overlapping domains showed a substantial agreement (Table 3). The median Gwet's AC_1 was 0.69 (range 0.38-0.84). Highest agreement was found for items/questions related to study selection. Partially overlapping domains only showed a slight agreement and were all lower than the full overlapping domains.

3.5. Applicability

Mean time for scoring AMSTAR 2 was slightly higher than that for ROBIS (18 min ± 7 min vs. 16 min ± 7 min, range 12 min to 22 for AMSTAR and 7 to 21 min for ROBIS). The type of review had no influence on scoring time for both tools (data not shown).

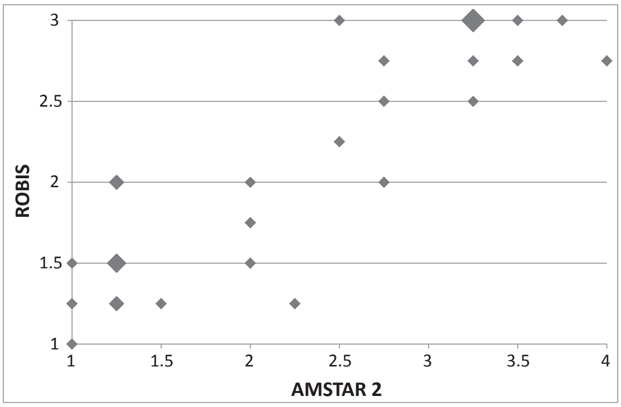

Fig. 1. Comparison of mean overall ratings for AMSTAR 2 and ROBIS. Each diamond represents one SR, while the thickness of the diamonds represents their number of occurrence, the thickest diamond (3.25/3) is n = 4.

Table 1. Fleiss kappa (95% CI) for four raters for AMSTAR 2

1	Did the research questions and inclusion criteria for the review include the components of PICO?	0.28 (0.05 to 0.52)
2	Did the report of the review contain an explicit statement that the review methods were established before conduction of the review and did the report justify any significant deviations?	1 (1 to 1)
3	Did the review authors explain their selection of the study designs for inclusion in the review?	0.65 (0.43 to 0.87)
4	Did the authors of this review use a comprehensive literature search strategy?	0.24 (0.09 to 0.39)
5	Did the review authors perform study selection in duplicate?	0.33 (−0.01 to 0.68)
6	Did the review authors perform data extraction in duplicate?	0.74 (0.50 to 0.98)
7	Did the review authors provide a list of excluded studies and justify the exclusions?	0.54 (0.37 to 0.70)
8	Did the review authors describe the included studies in adequate detail?	0.09 (−0.06 to 0.23)
9a	Did the review authors use a satisfactory technique for assessing the risk of bias in individual studies that were included in the review? RCTs	0.42 (0.23 to 0.62)
9b	Did the review authors use a satisfactory technique for assessing the risk of bias in individual studies that were included in the review? NRSI	0.31 (0.16 to 0.45)
10	Did the review authors report on the sources of funding for the studies included in the review?	0.70 (0.43 to 0.98)
11a	If meta-analysis was performed, did the review authors use appropriate methods for statistical combination of results? RCTs	0.61 (0.43 to 0.79)
11b	If meta-analysis was performed, did the review authors use appropriate methods for statistical combination of results? NRSI	0.53 (0.39 to 0.67)
12	If meta-analysis was performed, did the review authors assess the potential impact of risk of bias in individual studies on the results of the meta-analysis or other evidence synthesis?	0.62 (0.46 to 0.79)
13	Did the review authors account for risk of bias in individual studies when interpreting/discussing the results of the review?	0.35 (0.08 to 0.61)
14	Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	0.21 (0.01 to 0.42)
15	If they performed quantitative synthesis, did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	0.58 (0.41 to 0.75)
16	Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	0.46 (−0.13 to 1)
-	Overall confidence	0.30 (0.17 to 0.43)

Abbreviations: CI, confidence interval; AMSTAR, A Measurement Tool to Assess Systematic Reviews; NRSI, non-randomized studies of interventions; PICO, patient, intervention, comparison, outcome; RCTs, randomized controlled trials.

3.6. Influence of pair of reviewers

The pair of reviewers who have already worked together in a previous study [13] had a higher IRR than the other pair (Appendix E). For AMSTAR 2, the median IRR was Gwet's AC$_1$ 0.80 (range: 0.30 to 1) opposed to 0.58 (−0.09 to 1). The difference was biggest for item 8 and the overall confidence item. For both items, the 95% CI of the corresponding Gwet's AC$_1$ coefficients did not overlap between both pairs. A similar pattern was found for ROBIS, where IRRs were much higher for the same pair for all domains, except domain 4. The biggest difference was found for the overall domain, as indicated by Gwet's AC$_1$ of 0.47 (95% CI: 0.19 to 0.76) vs. 0.00 (95% CI: −0.28 to 0.28).

3.7. Cochrane reviews vs. non-cochrane reviews

For AMSTAR 2, a higher IRR was observed in nCRs than in CRs. The median Fleiss kappa was 0.19 (range: −0.11 to 1) for CRs compared to 0.51 (range: 0.04 to 1) for nCRs (Appendix F). The biggest differences were found in items 6 and 7, where the 95% CIs of the corresponding coefficients did not overlap and were higher for nCRs in all cases. However, a different pattern is seen when focusing on Gwet's AC$_1$ coefficient. According to that, IRR is higher in CRs than in nCRs, in particular for item 1 and item 7, where the 95% CIs do not overlap. This is very similar when looking at ROBIS, where no difference was observed in IRR between CRs and nCRs when measured with Fleiss Kappa, but with

Table 2. Fleiss kappa (95% CI) for four raters for ROBIS

Domain 1	Domain 2	Domain 3	Domain 4	Domain overall
0.42 (0.27 to 0.57)	0.21 (0.05 to 0.36)	0.23 (0.07 to 0.38)	0.20 (0.05 to 0.34)	0.28 (0.13 to 0.42)

Abbreviations: CI, confidence interval; ROBIS, risk of bias.

Table 3. AMSTAR 2/ROBIS matching domains

AMSTAR 2	ROBIS	Gwet's AC1
Domain 1: Study eligibility criteria		
1. Did the research questions and inclusion criteria for the review include the components of PICO?	1.3 Were eligibility criteria unambiguous?	0.65 (0.41-0.90)
2. Did the report of the review contain an explicit statement that the review methods were established before the conduct of the review and did the report justify any significant deviations from the protocol?	1.1 Did the review adhere to predefined objectives and eligibility criteria? (protocol)	0.48 (0.18-0.78)
Not considered	1.2 Were the eligibility criteria appropriate for the review question?	
3. Did the review authors explain their selection of the study designs for inclusion in the review?	*1.4 Were all restrictions in eligibility criteria based on study characteristics appropriate?*	*0.15 (-0.11-0.42)*
Not considered	1.5 Were any restrictions in eligibility criteria based on sources of information appropriate?	
Domain 2: Identification and selection of studies		
4. Did the review authors use a comprehensive literature search strategy?	2.1 Did the search include an appropriate range of databases/electronic sources for published and unpublished reports? 2.2 Were methods additional to database searching used to identify relevant reports? 2.3 Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible? 2.4 Were restrictions based on date, publication format, or language appropriate?[a]	0.38 (0.10-0.65)
5. Did the review authors perform study selection in duplicate?	2.5 Were efforts made to minimize error in selection of studies?	0.84 (0.69-1)
6. Did the review authors perform data extraction in duplicate?	3.1 Were efforts made to minimize error in data collection? 3.5 Were efforts made to minimize error in risk of bias assessment?[b]	0.75 (0.55-0.95)
Domain 3: Data collection and study appraisal		
7. Did the review authors provide a list of excluded studies and justify the exclusions?	Not considered	
8. Did the review authors describe the included studies in adequate detail?	3.2 Were sufficient study characteristics available for both review authors and readers to be able to interpret the results?	0.78 (0.59-0.96)
Not considered	3.3 Were all relevant study results collected for use in the synthesis?	
9. Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?[c]	3.4 Was risk of bias (or methodological quality) formally assessed using appropriate criteria?	0.54 (0.28-0.80)
10. Did the review authors report on the sources of funding for the studies included in the review?	Not considered	
Domain 4: Synthesis and findings		
Not considered	4.1 Did the synthesis include all studies that it should?	
Not considered	4.2 Were all predefined analyses reported or departures explained?	
11. If meta-analysis was performed, did the review authors use appropriate methods for statistical combination of results?[c]	4.3 Was the synthesis appropriate, given the nature and similarity in the research questions, study designs, and outcomes across included studies? 4.4 Was between-study variation (heterogeneity) minimal or addressed in the synthesis?[d]	0.76 (0.52-1)
12. If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis? 13. Did the review authors account for RoB in	4.6 Were biases in primary studies minimal or addressed in the synthesis?	0.73 (0.52-0.94)

(*Continued*)

Table 3. Continued

AMSTAR 2	ROBIS	Gwet's AC1
individual studies when interpreting/ discussing the results of the review?[e]		
14. Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	4.4 Was between-study variation (heterogeneity) minimal or addressed in the synthesis	0.64 (0.40-0.89)
15. If they performed quantitative synthesis, did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	*4.5 Were the findings robust, for example, as demonstrated through funnel plot or sensitivity analyses?*	*0.05 (-0.23-0.34)*
16. Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	Not considered	

For AMSTAR 2, items rated with ''partial yes'' or ''no meta-analysis conducted'' were rated ''yes'' in all cases except item 9, where ''partial yes'' was regarded as ''no.'' For ROBIS, we combined the ''yes'' and ''probably yes'' and the ''probably no'' and ''no'' answers to signaling questions, as recommended by the developers of ROBIS [10]. If raters' judgments were evenly distributed, we coded this as a missing. Partial overlap is marked in italics.

[a] If 2.1 AND 2.2 AND 2.3 AND 2.4 = yes, then = YES, else = NO.

[b] If 3.1 AND 3.5 = yes, then = YES, else = NO.

[c] RCTs and NRSI = yes, then YES, else = NO.

[d] If 4.3 and 4.4 = yes, then = YES, else = NO.

[e] If 12 and 13 = yes, then YES, else = NO.

Gwet's AC_1 coefficient. The biggest difference was found for domains 1 and 2, where IRR was much higher in CRs than in nCRs as indicated by nonoverlapping CIs.

4. Discussion

This is the first study to compare AMSTAR 2 with ROBIS with respect to SRs that included both RCTs and non-RCTs. Our study suggests that reliability was slightly higher for AMSTAR 2 than for ROBIS, although there were also problematic items in AMSTAR 2 in terms of IRR. The IRR for the overall assessment for both tools was fair. Results for IRR were dependent on the pair of reviewers, as raters who worked together in similar projects in the past had higher IRR. There was also a high correlation between both tools, suggesting validity. This was also supported when matching corresponding items/questions of AMSTAR 2 and ROBIS. Applicability was found to be satisfactory for both tools. AMSTAR 2 is an improvement compared to AMSTAR.

To the best of our knowledge, this is the first validation study for AMSTAR 2. Thus, the only study we can compare our results with is the source publication [23]. In general, our IRR was slightly lower, although there are also items where our IRR was higher than in the three pairs of raters in the AMSTAR 2 development study. In contrast to us, they did not face difficulties for item 8 and item 14. For items 9 and 11, the IRR for risk of bias judgments for RCTs were similar to those for non-RCTs. This is in congruence with testing the tool by its developers [8]. With respect to IRR for ROBIS, other studies have shown higher IRR [13,24,25] or similar IRR [12,26]. All comparison studies only dealt with SRs of RCTs, what might at least in part serve as an explanation for our lower IRR values.

We have obtained a very high correlation between the overall ratings of AMSTAR 2 and ROBIS. It has to be kept in mind that these two tools measure a different concept. AMSTAR 2 focuses on methodological quality, whereas ROBIS focuses on risk of bias. Nevertheless, both concepts are strongly related to each other [4]. Our correlation was much higher than that others have observed for ROBIS and AMSTAR [12,13]. This was probably due to the fact that AMSTAR 2 is much more detailed than AMSTAR, with more specific questions addressing each phase of the review process, and has more domains matching with ROBIS domains compared to AMSTAR.

There was also a strong agreement when matching AMSTAR 2 items with their corresponding ROBIS signaling questions and vice versa. This should only be treated as an explanatory analysis. It does not fully demonstrate concurrent validity as both tools cannot be applied at the same time. Therefore, the order in that the ratings were performed might have an influence on this finding.

We have observed that scoring time was slightly higher for AMSTAR 2 than for ROBIS. However, it must be taken into account that all reviewers always started with AMSTAR 2, and reading time is therefore included in the time taken for AMSTAR 2. The developers of AMSTAR 2 reported that it took them between 15 and 32 minutes to apply AMSTAR 2 [8]. This range excludes the reading time. With 18 minutes on average, including reading time, we were much faster.

Raters found AMSTAR 2 easier to apply, with questions more clear, simple and specific; in addition, the AMSTAR 2 guidance was found to be clearer and simpler than the

ROBIS guidance; this would probably facilitate its use also by nonexperienced reviewers. Moreover, for a careful application of ROBIS, a certain degree of knowledge of subject matter of the review is necessary, while this expertise is not necessary to answer AMSTAR 2 questions, which is only focused on the rigor of the methods while ROBIS focuses more on the results section, whether a given procedure (or its lack) had an impact on the validity of the review findings.

An important finding of our study is the higher IRR for reviewers who have worked together on earlier occasions, although only for ROBIS. The IRR among the reviewers who have not worked together was slight in many items and domains. However, it is important to keep in mind that we did not perform any calibration exercise before our study. A calibration exercise might have resulted in higher IRR for the naïve pair, whereas the other pair can be regarded as being calibrated having worked together on a past project regarding ROBIS. The same tendency can be observed for AMSTAR 2, although it is not that clear as for ROBIS, as the naïve pair had higher IRR than the experienced pair in some items. Interestingly, differences among pair of reviewers were also spotted by the developers of AMSTAR 2. When reviewers decide to use either ROBIS or AMSTAR 2, we suggest performing calibration exercises before.

Our results could also indicate the need for further improvement of these two tools. Some of the items are very specific, such as second item of AMSTAR 2, which asks whether SR methods were established a priori. On the contrary, many items of AMSTAR 2 are asking for rating of "comprehensive literature search strategy," "adequate details," "satisfactory technique," and "satisfactory explanation"; these items are more prone to subjective biases and opinions of individual reviewers about what does it mean comprehensiveness, adequacy, and satisfaction. Therefore, lower IRR on those items might also be reflection of reviewers' subjective opinions about what constitutes a proper SR methodology. This is strong argument in favor of performing a calibration exercise before using these tools.

4.1. Limitations

Our sample of 30 SRs was derived from two different sources, whereas one of them focused on anesthesiology and pain. Our results based on that sample for nCRs might be biased by this fact as quality of SRs can differ across different fields of medicine. Another limitation is that no consensus procedure took place to obtain final ratings. Instead, we relied on means or majority of the raters' judgments.

5. Conclusion

Both AMSTAR 2 and ROBIS can be applied to SRs including both RCTs and non-RCTs. Measurement properties of ROBIS seemed not to be much different when comparing with other studies that include only SRs of RCTs. This needs to be further investigated in particular for the newer AMSTAR 2, and in addition, in SRs of non-RCTs. AMSTAR 2 is easier to apply than ROBIS.

Acknowledgments

The authors thank Ms Tanja Rombey for data verification.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.12.004.

References

[1] 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.

[2] 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.

[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] 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.

[5] 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.

[6] Faggion CM Jr. Critical appraisal of AMSTAR: challenges, limitations, and potential solutions from the perspective of an assessor. BMC Med Res Methodol 2015;15:63.

[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] Norris SL, Atkins D. Challenges in using nonrandomized studies in systematic reviews of treatment interventions. Ann Intern Med 2005;142:1112–9.

[10] Whiting P, Savović 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.

[11] Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Version 5.1.0. The Cochrane Collaboration; 2011. Available from www.handbook.cochrane.org; http://handbook-5-1.cochrane.org.

[12] 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.

[13] 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.

[14] Petticrew M, Wilson P, Wright K, Song F. Quality of cochrane reviews. Quality of cochrane reviews is better than that of non-cochrane reviews. BMJ 2002;324:545.

[15] Windsor B, Popovich I, Jordan V, Showell M, Shea B, Farquhar C. Methodological quality of systematic reviews in subfertility: a comparison of Cochrane and non-cochrane systematic reviews in assisted reproductive technologies. Hum Reprod 2012;27:3460−6.

[16] Fleming PS, Seehra J, Polychronopoulou A, Fedorowicz Z, Pandis N. A PRISMA assessment of the reporting quality of systematic reviews in orthodontics. Angle Orthod 2013;83(1):158−63.

[17] Biocic M, Fidahic M, Cikes K, Puljak L. Information sources used in systematic reviews of randomized controlled trials in the field of anesthesiology and pain 2018: [unpublished].

[18] Polus S, Pieper D, Burns J, Fretheim A, Ramsay C, Higgins JPT, et al. Heterogeneity in application, design, and analysis characteristics was found for controlled before-after and interrupted time series studies included in cochrane reviews. J Clin Epidemiol 2017;91:56−69.

[19] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378.

[20] Gwet KL. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC; 2014. Available from www.handbook.cochrane.org; http://handbook-5-1.cochrane.org.

[21] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.

[22] 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.

[23] da Costa BR, Beckett B, Diaz A, Resta NM, Johnston BC, Egger M, et al. Effect of standardized training on the reliability of the Cochrane risk of bias assessment tool: a prospective study. Syst Rev 2017;6(1):44.

[24] 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.

[25] 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.

[26] 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.