Reproducibility of risk of bias assessment using the SYRCLE tool with and without the assistance of a customized flowchart

Sofija Vojvodic, María Arroyo Araujo, Torsten Rackoll
QUEST Center for Responsible Research, Berlin Institute of Health, Charité University of Medicine

Introduction / Aim

A systematic review (SR) is only as reliable as the studies it includes, making the assessment of their methodological quality a crucial step in synthesizing evidence. With the growing number of SRs focused on animal studies, it became evident that animal studies differ in key aspects from human clinical trials, which are more commonly synthesized in SRs. As a result, specific guidance was needed to assess the internal validity of animal studies. In response, the SYRCLE Risk of Bias (RoB) tool was developed by Hooijmans et al. (2014)[1], adapting the Cochrane RoB tool originally designed for randomized controlled trials (RCTs).

The SYRCLE RoB tool evaluates potential biases—such as selection, performance, detection, attrition, reporting, and other sources of bias—across 10 items tailored specifically to animal intervention studies. It includes signalling questions that guide the user in determining whether each domain is at low, high, or unclear risk of bias. As with other steps in conducting an SR, the gold standard for quality assessment is to have it performed by at least two independent reviewers. This minimizes errors and reduces bias. However, a significant challenge in applying the SYRCLE RoB tool is the frequent lack of detailed and transparent reporting of the definitions of the tool items dependent to specific research questions, which can make interpretations of some signalling questions subjective.

To address this issue and promote more standardized use, we developed a customized flowchart to guide the application of the SYRCLE RoB tool. This flowchart breaks down each item into a series of guiding questions, allowing reviewers to assign scores more systematically and consistently. To test whether our flowchart would lead to more consistent ratings, we sought to reproduce the risk of bias assessments from a published SR by Ripley et al. (2021)[2] using the flowchart, and then compare these assessments against those made without the flowchart. Thus, we had two aims in this reproducibility challenge:

1. Reproduce the Risk of Bias Assessments: We aimed to replicate the risk of bias assessments from Ripley et al. (2021), where the SYRCLE RoB tool was used to evaluate 76 studies on remote ischemic conditioning (RIC) as an intervention in rodent models of transient cerebral ischemia. In the original study, two independent reviewers conducted the risk of bias assessments, and discrepancies were resolved by a third reviewer. In our study, we applied the same method and definitions to half of the included studies (k=36) to determine if the original risk of bias assessments could be successfully replicated.
2. Evaluate the Impact of the RoB Flowchart on Reviewer Agreement: We sought to evaluate whether using the RoB flowchart improves inter-rater reliability compared to using only the definitions for each item. The flowchart was applied to the remaining half of the included studies. Our goal was to determine if the flowchart improves the level of agreement between reviewers, making the RoB assessment more robust, replicable, and time-efficient.

Methods

All the materials used in the reproducibility challenge and detailed descriptions of risk of bias methods are publicly available in OSF.

We requested the list of 72 studies included in the systematic review from the authors and uploaded this data into the SyRF platform[3] for assessment. The studies were then randomly assigned to be assessed by one of two scoring methods using the Microsoft Excel function "randbetween(1,2)," ensuring equal distribution between methods. Each paper was independently assessed by two reviewers (MA and SV), with discrepancies resolved through discussion. If disagreements persisted, a third reviewer (TR) reconciled them. The two scoring methods, both based on SYRCLE's Risk of Bias (RoB) tool, were as follows:

1. List of Definitions Used in the Original Paper: We contacted the authors of the original study to clarify their interpretation of each item in the SYRCLE tool, obtaining definitions for scoring each item as high, low, and unclear risk of bias. This was used as the basis for assessing half of the studies.
2. Risk of Bias Flowchart: This method used a decision tree to guide the scoring of each item in the SYRCLE tool. For certain items that required study-specific definitions, we referred to the original authors' definitions. For example, item 2, concerning baseline characteristics, was assessed according to the definitions from the original SR. In cases in which the implementation of the SYRCLE item was not aligned and thus non-comparable between the list and the flowchart, the item was still scored but flagged to be taken out of the analysis for aim 2. This was the case for items 9 and 10.

To ensure consistency, both reviewers had prior experience with the flowchart and first scored the papers assigned to the original list method before moving on to the flowchart assessment. In addition to the list of included papers and the SYRCLE RoB tool implementation used for the original RoB assessment, the authors also shared translations they had used for non-English language papers.

Results and evaluation of reproducibility

The complete dataset consisted of 71 studies; one non-English study belonging to method 2 was excluded because its translation was uninterpretable. The RoB assessment of 2 studies belonging to method 2 were lost so the inter-rater analysis was based on the 36 studies for method 1 and 33 studies for method 2.

Two reviewers could not agree on the assessments for 37 studies, and due to time constraints, only 20 of these were reconciled by a third reviewer. Since the original study authors also resolved disagreements with a third reviewer, we focus on replication attempts for single study assessments that were reconciled in the same way (see Appendix 1). All reconciled assessments adhered to the original authors' definitions of the SYRCLE items. However, most risk of bias items were not fully reproduced. Exceptions were performance bias (random housing) and detection bias (random outcome assessment), which were consistently rated as unclear due to insufficient reporting. The assessment of blinded outcome assessments (item 7) was successfully reproduced for all reconciled papers.

For selection bias, including sequence generation and baseline characteristics, ratings were generally unclear, with high and low ratings observed in two studies (Hoda et al. 2014, Chen et al. 2018), respectively. Allocation concealment was consistently rated as unclear during the reproducibility challenge, contrasting with the original SR, where four studies scored high and one scored low on this item. Risk of inappropriate blinding of experimenters was also rated as unclear in the reproducibility challenge but was mostly rated high in the original SR (11/20). The pattern of ratings for the risk of selective reporting was similar, with most studies rated as high risk of selective reporting remaining consistent between the reproducibility challenge and the original SR (11/15). At the single study level, assessments were fully reproduced for only one study (Kitagawa et al. 2018).

Overall, we reproduced assessments for 3 out of 10 items, with two items consistently rated as unclear across all studies. Selective reporting was partially reproduced.

The inter-rater reliability was consistent with our expectations, most disagreements (29/37) occurred in studies assessed using method 1 (without the flowchart), suggesting that the use of the standardized flowchart leads to more inter-rater consistency. We further compared the congruence of ratings for each bias domain, both with and without the flowchart (Appendix 1, Table 3). Across domains, inter-rater agreement was higher when assessments were performed using the flowchart, reinforcing its value in promoting more systematic and reliable evaluations.

Discussion / Conclusion

In attempting to reproduce the RoB assessment of a preclinical systematic review, we encountered several challenges that highlight broader issues in the reproducibility of SRs of animal studies. Our experience offers important insights into both the difficulties and potential solutions for improving the robustness and reproducibility of preclinical SRs.

One of the most significant challenges we faced was the time-consuming nature of the RoB assessment process. Even though we initially underestimated the time required - assuming that we wouldn't need to read entire papers, only focus on specific elements related to bias - it still took over 80 person-hours to assess 72 studies and discuss/reconcile disagreements. This experience underscores a broader issue in SRs: study quality assessments are resource-intensive. Given that RoB assessments are critical for ensuring the reliability of systematic reviews, there is a clear need to explore ways to make these assessments more efficient. One potential avenue is the development of automation tools. The integration of machine learning or artificial intelligence to assist with the identification of potential biases and extraction of relevant data could drastically reduce the time required for these tasks. Automation could serve as a tool for initial screening, allowing human reviewers to focus on more nuanced aspects of the assessment.

Another challenge we encountered was related to the use of different data management tools during the RoB assessment. While we used SyRF for our assessment, the original authors used DistillerSR, and this mismatch led to various complications. The differences in data management tools introduced discrepancies in the level of detail stored, complicating the comparison of our results with the original study. For instance, we were unable to access the original single-reviewer (non-reconciled) assessments for direct comparison with our single-reviewer assessments, which limited the depth of our analysis. Furthermore, differences in data formatting between these platforms made the data cleaning process more time-consuming. We also noticed inconsistencies in reference metadata, such as different publication years for the same studies (e.g., Xia et al.; Liu et al.), likely arising from the use of distinct software.

Another issue we faced was the incomplete or unclear documentation of the original authors' definitions of the SYRCLE items. This lack of clarity made it difficult to interpret certain aspects of the RoB tool during our replication process. Although the original authors were responsive and helpful, they could not recall all the details of their methods due to the time elapsed since the original review. This situation highlights a critical problem for reproducibility: without precise documentation of how specific tools and criteria were applied in a given study, replicating these efforts becomes extremely challenging. To overcome this, it is crucial that systematic reviews document their methods, including the interpretation of RoB items, or any deviations from the original tool, in thorough detail.

Our use of a customized flowchart to guide the SYRCLE RoB tool application provides an important insight into how the reproducibility of RoB assessments can be improved. The flowchart operationalized each item by breaking it down into specific guiding questions, which made the assessment process more systematic and less prone to subjective interpretation. Our findings show that using the flowchart led to fewer discrepancies between reviewers compared to assessments based solely on original authors' definitions. This suggests that structured tools like flowcharts can help standardize the assessment process and reduce inconsistencies. Given the challenges we experienced with the subjective interpretation of certain RoB items, we believe that the use of flowcharts or other structured decision-making tools should become a standard part of RoB assessments in preclinical SRs.

Lastly, our experience underscores the importance of open science and transparency in the SR process. Throughout our attempt to replicate the original RoB assessments, access to clear documentation, data, and methods was crucial. SR reporting guidelines such as PRISMA emphasize the need for clear, comprehensive documentation of all aspects of the review process, with the goal to ensure that any SR can be replicated or updated in the future. However, our experience highlights that while this expectation exists in theory, in practice, we still face challenges in achieving true transparency and replicability. The original authors' willingness to share their materials with us greatly facilitated our work, but gaps in documentation still posed challenges. Going forward, it is essential that SR authors make their methods and data better documented and even more FAIR (findable-accessible-interoperable-reusable)[4].

In summary, the challenges we encountered during this replication attempt—ranging from time constraints to data management issues and incomplete documentation—underscore the complexities of ensuring reproducibility in SRs and meta-analyses of preclinical research. However, the use of structured tools like flowcharts, better documentation, and increased transparency are all strategies that can mitigate these issues and improve the robustness of future reviews.

References

1. Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. BMC Med Res Methodol. 2014 Mar 26;14:43. doi: 10.1186/1471-2288-14-43. PMID: 24667063; PMCID: PMC4230647.
2. Ripley AJ, Jeffers MS, McDonald MW, Montroy J, Dykes A, Fergusson DA, Silasi G, Lalu MM, Corbett D. Neuroprotection by Remote Ischemic Conditioning in Rodent Models of Focal Ischemia: a Systematic Review and Meta-Analysis. Transl Stroke Res. 2021 Jun;12(3):461-473. doi: 10.1007/s12975-020-00882-1. Epub 2021 Jan 6. PMID: 33405011.
3. Bahor Z, Liao J, Currie G, Ayder C, Macleod M, McCann SK, Bannach-Brown A, Wever K, Soliman N, Wang Q, Doran-Constant L, Young L, Sena ES, Sena C. Development and uptake of an online systematic review platform: the early years of the CAMARADES Systematic Review Facility (SyRF). BMJ Open Sci. 2021 Mar 30;5(1):e100103. doi: 10.1136/bmjos-2020-100103. PMID: 35047698; PMCID: PMC8647599.
4. Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., … Schultes, E. (2020). FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence, 2(1–2), 10–29. https://doi.org/10.1162/dint_r_00024

Acknowledgments

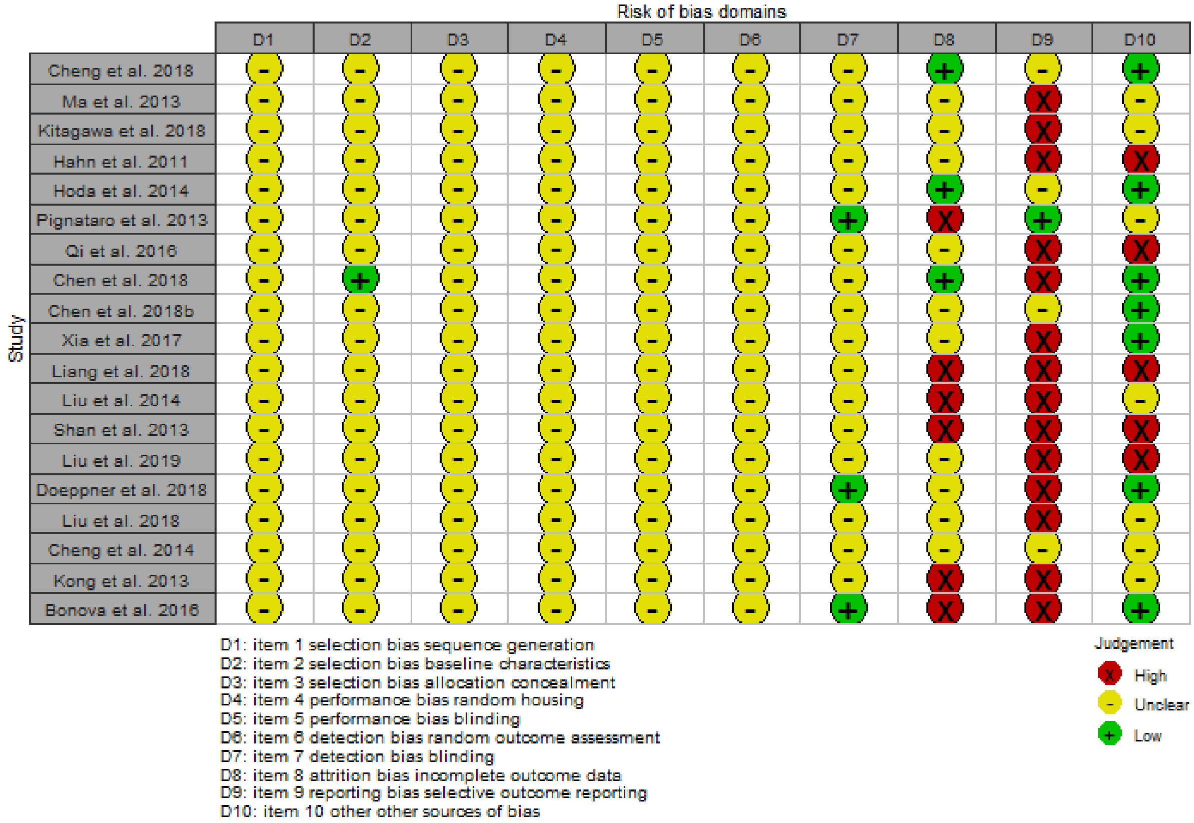
The writing of this report was facilitated using ChatGPT3.0.

Appendix 1 – Comparison of risk of bias assessments in reproducibility challenge and original study

Table 1. Adapted Supplement table 7 from Ripley et al. 2021 to include only studies for which the risk of bias assessments were reconciled in the reproducibility challenge

References	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Cheng 2018	🟨	🟨	🟨	🟨	🟥	🟨	🟨	🟥	🟥	🟨
Ma 2013	🟨	🟨	🟥	🟨	🟥	🟨	🟨	🟨	🟥	🟩
Kitagawa 2018	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟥	🟨
Hahn 2011	🟨	🟨	🟨	🟨	🟥	🟨	🟨	🟨	🟥	🟨
Hoda 2013	🟥	🟨	🟩	🟨	🟩	🟨	🟨	🟩	🟩	🟨
Pignataro 2013	🟨	🟨	🟨	🟨	🟥	🟨	🟩	🟥	🟩	🟩
Qi 2016	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟥	🟥	🟩
Chen 2018	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟥	🟥	🟩
Chen 2018	🟨	🟨	🟥	🟨	🟥	🟨	🟨	🟥	🟥	🟨
Xia 2016	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟥	🟥	🟨
Liang 2018	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟥	🟥	🟨
Liu 2014	🟨	🟨	🟨	🟨	🟥	🟨	🟨	🟨	🟥	🟨
Shan 2013	🟨	🟨	🟥	🟨	🟥	🟨	🟨	🟥	🟥	🟨
Liu 2018	🟨	🟨	🟨	🟨	🟥	🟨	🟨	🟥	🟥	🟥
Doeppner 2018	🟨	🟨	🟨	🟨	🟨	🟨	🟩	🟩	🟩	🟨
Liu 2018	🟨	🟨	🟨	🟨	🟥	🟨	🟨	🟨	🟥	🟨
Cheng 2014	🟨	🟨	🟥	🟨	🟥	🟨	🟨	🟨	🟥	🟥
Kong 2013	🟨	🟨	🟨	🟨	🟥	🟨	🟨	🟥	🟩	🟥
Bonova 2016	🟨	🟨	🟨	🟨	🟨	🟨	🟩	🟥	🟥	🟨

Table 2. Reconciled risk of bias assessments in the reproducibility challenge



Risk of bias domains

Study	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Cheng et al. 2018	-	-	-	-	-	-	-	+	-	+
Ma et al. 2013	-	-	-	-	-	-	-	-	X	-
Kitagawa et al. 2018	-	-	-	-	-	-	-	-	X	-
Hahn et al. 2011	-	-	-	-	-	-	-	-	X	X
Hoda et al. 2014	-	-	-	-	-	-	-	+	-	+
Pignataro et al. 2013	-	-	-	-	-	-	+	X	+	-
Qi et al. 2016	-	-	-	-	-	-	-	-	X	X
Chen et al. 2018	-	+	-	-	-	-	-	+	X	+
Chen et al. 2018b	-	-	-	-	-	-	-	-	-	+
Xia et al. 2017	-	-	-	-	-	-	-	-	X	+
Liang et al. 2018	-	-	-	-	-	-	-	X	X	X
Liu et al. 2014	-	-	-	-	-	-	-	X	X	-
Shan et al. 2013	-	-	-	-	-	-	-	X	X	X
Liu et al. 2019	-	-	-	-	-	-	-	-	X	X
Doeppner et al. 2018	-	-	-	-	-	-	+	-	X	+
Liu et al. 2018	-	-	-	-	-	-	-	-	X	-
Cheng et al. 2014	-	-	-	-	-	-	-	-	-	-
Kong et al. 2013	-	-	-	-	-	-	-	X	X	-
Bonova et al. 2016	-	-	-	-	-	-	+	X	X	+

D1: item 1 selection bias sequence generation
D2: item 2 selection bias baseline characteristics
D3: item 3 selection bias allocation concealment
D4: item 4 performance bias random housing
D5: item 5 performance bias blinding
D6: item 6 detection bias random outcome assessment
D7: item 7 detection bias blinding
D8: item 8 attrition bias incomplete outcome data
D9: item 9 reporting bias selective outcome reporting
D10: item 10 other other sources of bias

Judgement

X High
- Unclear
+ Low

Table3. Comparison of inter-rater consistency by method of RoB assessment

RoB Item	Count agreements-Method 1	% agreement Method 1	Count agreements-Method 2	% agreement Method 2
1	36	100	33	100
2	33	91	33	100
3	36	100	33	100
4	36	100	33	100
5	36	100	33	100
6	36	100	33	100
7	36	100	32	96
8	34	94	25	75

*Note – Items 9 and 10 were not comparable between the two methods therefore they were left out.