

MASTER scale for methodological quality assessment: Reliability assessment and update

In evidence synthesis of analytical studies, methodological quality (mQ) assessment is necessary to determine the extent to which internal validity safeguards are implemented in the included studies against a list of selected safeguards in an assessment tool. Such mQ tools consist of internal validity safeguards that are checked against those put in place by researchers when they undertake research to guard against systematic error in the design, conduct, and analysis of a study.¹ However, consistency or agreement among the individuals undertaking an assessment of implemented safeguards in published research against those listed in a mQ tool needs to be documented to ensure that the tool is reliable. Therefore mQ tools need to have their inter-rater reliability tested in order to ensure the consistency of their use in research.²

Many existing tools are available to assess mQ or risk of bias (RoB) specific to a study design, which leads to a lack of comparability across studies of different designs when using different tools and assessment results which, as a whole, may lack meaning. For example, Cochrane's Risk of Bias (RoB2) tool is used to assess the RoB in RCTs while non-randomized trials are assessed using the ROBINS-I tool. It is difficult to compare these scales to one another, and hence, there is a need for a unified scale that is not confined to one study design. The MASTER scale was developed to overcome some of these issues by providing a comprehensive list of methodological safeguards across analytic study designs that allow for comparative assessment between these studies. It uses an assessment approach that takes the reviewer all the way from mQ assessment through to an ability to make use of this for bias adjustment.^{3,4} A drawback for reviewers using the MASTER scale is that there is a lack of information regarding its reliability, with no studies conducted to assess this metric.

The degree to which studies maintain their relative position in a list over repeated measurements is referred to as reliability.⁵ For example, when assessing the reliability of a tool such as the MASTER scale, it would be considered reliable if you see that studies which scored well on the tool by the first rater also scored well on subsequent assessments by different raters.^{5,6} The scoring system for this scale has been discussed previously.⁷ Such consistency across the individuals undertaking mQ assessment needs to be established to ensure that the tool is reliable across different raters. Researchers trained in clinical epidemi-

ology were chosen for this study so that they could also examine the scale item wordings to remove ambiguity and improve the readability and applicability of the wording. This study therefore serves the dual purpose of evaluating the reliability of the MASTER scale across raters and examine the scale wording to see if the tool needs to be updated for clarity and readability.

As shown in Table S1, there were 11 studies⁸⁻¹⁸ chosen for assessment that contained a total of 1344 patients conducted using different study designs comparing normal saline with ringers' lactate in the treatment of acute pancreatitis. Five^{9-11,13,18} of the 11 studies were randomized-controlled trials including 299 patients, three^{8,12,14} were cohort studies including 433 patients, and three¹⁵⁻¹⁷ were abstracts with 612 patients and the designs reported within the abstracts were observational in one and possibly experimental in two. The highest mean quality safeguard count (Qi) across the raters was observed in the study by De-Madaria¹⁰ at 33.17 (SD 1.33). Conversely, the lowest mean Qi was reported in the study by Vasu De Van,¹⁷ an abstract based on a RCT of 50 patients, with a mean of 8.83 (SD 4.45). The highest mean for the relative rank was again found in the study by De-Madaria¹⁰ at 0.99 (SD 0.01), while the lowest relative rank was noted in the study by Vasu De Van¹⁷ at 0.27 (SD 0.12). Similarly, for the absolute ranks, the highest mean was observed in the study by De-Madaria¹⁰ at 1.17 (SD 0.41), and the lowest was in the study by Vasu De Van¹⁷ with a mean of 10.67 (SD 0.52). It should be noted that the study with the highest count always has a relative rank of 1 and this would decrease as the study rank gets lower⁷. On the other hand, absolute ranks are also highest at 1 but increase as ranks get lower.

Figure S1 illustrates how the six raters evaluated one of the eleven studies. The graph shows the overall safeguard count that each rater assigned as well as a breakdown analysis of the overall count that demonstrates the amount of the total count that was contributed by each standard. The results indicate high internal consistency and reliability for all three measures as shown in Table S2. The total safeguard count (Qi) and relative ranks yielded an ICC of 0.90 (95% CI: 0.79-0.97) and 0.90 (95% CI: 0.80-0.97), respectively indicating excellent level of agreement between raters. The absolute ranking measure had the highest level of agreement, with an ICC of 0.93 (95% CI: 0.86-0.98). Overall, the results suggest that there is low disparity of the

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Journal of Evidence-Based Medicine* published by Chinese Cochrane Center, West China Hospital of Sichuan University and John Wiley & Sons Australia, Ltd.

TABLE 1 Updated MASTER scale V1.01.

Standard	MASTER scale V1.01	Previous wording
1. Equal recruitment	1. Data collected after the start of the study (unrelated to eligibility) were not used to exclude participants or to select them into the analysis	1. Data collected after the start of the study was not used to exclude participants or to select them into the analysis
	2. Participants in all comparison groups met the same eligibility requirements and were from the same population and timeframe	Same
	3. Start of follow-up, determination of eligibility and assignment to comparison groups were synchronized or this was an experimental study	3. Determination of eligibility and assignment to treatment group/exposure strategy was synchronized
	4. None of the eligibility criteria were common effects (colliders) of exposure and outcome	4. None of the eligibility criteria were common effects of exposure and outcome
2. Equal retention	5. Any attrition in prospective studies is less than 20% of total participant numbers (safeguard absent in retrospective studies)	5. Any attrition (or exclusions after entry) is less than 20% of total participant numbers
	6. Missing covariates (not exposure or outcome) were in less than 20% of participants	6. Missing data is less than 20%
	7. Analysis accounted for missing data	Same
	8. Exposure or treatment definition deviations from protocol were unlikely to impact the final outcome	8. Exposure variations/treatment deviations were less than 20%
	9. Changes to exposure/intervention or withdrawals after the start of the study were addressed by analysis	9. Variations in exposure or withdrawals after the start of the study were addressed by analysis
3. Equal ascertainment	10. Procedures for data collection of covariates were reliable and the same for all participants	Same
	11. The outcome was objective and/ or reliably measured	Same
	12. Exposures/ interventions were objectively and/ or reliably measured	Same
	13. Outcome assessor(s) were blinded	Same
	14. Participants were blinded	Same
	15. Caregivers were blinded	Same
	16. Analyst(s) were blinded	Same
4. Equal implementation	17. Care was delivered equally to all participants	Same
	18. Cointerventions that could impact the outcome were comparable between groups or avoided	Same
	19. Control and active interventions/ exposures were sufficiently distinct	Same
	20. Exposure/intervention definition was consistently applied to all participants	Same
	21. Outcome definition was consistently applied to all participants	Same
	22. The follow-up time period is similar across patients and between groups or the analyses adjust for different lengths of follow-up of patients	22. The time period between exposure and outcome is similar across patients and between groups or the analyses adjust for different lengths of follow-up of patients
5. Equal prognosis	23. Design and/or analysis strategies were in place that addressed potential confounding	Same
	24. Method of selection of confounders ensured that they were not common effects (colliders) of exposure and outcome or this is a randomized trial	24. Key confounders addressed through design or analysis were not common effects of exposure and outcome
	25. There were no major prognostic differences across groups or this was addressed through analysis	25. Key baseline characteristics / prognostic indicators for the study were comparable across groups
	26. Participants were randomly allocated to groups with an adequate randomization process reported	26. Participants were randomly allocated to groups with adequate randomization process
	27. Allocation procedure was adequately concealed	Same
	28. Conflict of interests were declared and absent	Same

(Continues)

TABLE 1 (Continued)

Standard	MASTER scale V1.01	Previous wording
6. Sufficient analysis	29. Analytic method was justified by study design or data requirements	Same
	30. Computation errors or contradictions were absent	Same
	31. There was no discernible data dredging or selective reporting of the outcomes	Same
7. Temporal precedence	32. All subjects were selected prior to intervention/ exposure and evaluated prospectively	Same
	33. Carry-over or refractory effects were avoided or considered in the design of the study or were not relevant	Same
	34. The intervention/ exposure period was long enough to have influenced the study outcome	Same
	35. Dose of intervention/ exposure was sufficient to influence the outcome	Same
	36. Length of follow-up was not too long or too short in relation to the outcome assessment	Same

overall raters' evaluation of an aggregate assessment using the MASTER scale.

When looking across each of the individual standards, there was strong interrater reliability (Table S3). For instance, standard 3 (ICC 0.89, 95% CI 0.78–0.96) made the biggest contribution to the overall reliability across raters for this study. However, for all six raters, standards 1 (ICC 0.61, 95% CI 0.36–0.84), 4 (ICC 0.62, 95% CI 0.38–0.85), 6 (ICC 0.66, 95% CI 0.43–0.87), and 7 (ICC 0.61, 95% CI 0.36–0.84) had the most room for improvement in terms of reliability in this study (Table S3). Overall, these results suggest that there is moderate to excellent agreement among the raters within the MASTER scale standards.

Table 1 depicts the updated MASTER scale depicting areas of the MASTER scale where recommended changes to the wordings of safeguards were made. Overall, 26 safeguards had suggestions raised within the following four standards of the MASTER scale “Equal recruitment,” “Equal retention,” “Equal implementation” and “Equal prognosis.” However, the following standards, “Equal ascertainment,” “Sufficient analysis,” and “Temporal precedence”, had no suggestions raised. We present this version of the MASTER scale for future use as version 1.01.

In conclusion, the MASTER scale (updated V1.01, Table 1) appears to be a reliable unified (across analytical study designs) tool for assessing individual studies in evidence syntheses. Our study has identified some areas where the wording of the scale could be improved, which would enhance its clarity and further increase its reliability. The main issues flagged were around the wording of the questions, and how they could be improved for interpretation and understanding, especially by those not experts in clinical epidemiology. The main limitation of using any scale, not just the MASTER scale, is the time and expertise required for generating the assessment. Other than this, the MASTER scale has no other significant limitations. This opens the door for further research in examining the reliability of the MASTER scale when assessed by other health students, clinical researchers, and other health care work-

ers. Overall, the findings of this study have significant implications for future use and wider adoption of the MASTER scale in evidence synthesis due its applicability to all types of study designs.

ACKNOWLEDGMENTS

This work was presented at the University of Warwick to the World Congress on Undergraduate Research 2023 (DATA theme).




Open Access funding provided by the Qatar National Library.

CONFLICT OF INTEREST STATEMENT

Authors JS and SD were responsible for the creation of the MASTER scale. There are no other interests to report.

FUNDING INFORMATION

We acknowledge the financial support provided by the College of Medicine at Qatar University, which enabled the successful completion of this research project.

Ashraf I. Ahmed¹
 Muhammad Zain Kaleem¹
 Amgad Mohamed Elshoeibi¹
 Abdalla Moustafa Elsayed¹ 
 Elhassan Mahmoud¹
 Yaman A. Khamis¹
 Luis Furuya-Kanamori² 
 Jennifer C. Stone³ 
 Suhail A. Doi¹

¹Department of Population Medicine, College of Medicine, QU Health, Qatar University, Doha, Qatar

²UQ Centre for Clinical Research, The University of Queensland, Herston, Queensland, Australia

³JBI, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, South Australia, Australia

Correspondence

Jennifer C. Stone, JBI, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, South Australia, Australia.

Email: j.stone@adelaide.edu.au

Suhail A. Doi, Department of Population Medicine, College of Medicine, Qatar University, Doha, Qatar.

Email: sdoi@qu.edu.qa

ORCID

Abdalla Moustafa Elsayed  <https://orcid.org/0000-0001-9683-0540>

Luis Furuya-Kanamori  <https://orcid.org/0000-0002-4337-9757>

Jennifer C. Stone  <https://orcid.org/0000-0002-3787-6175>

REFERENCES

1. Furuya-Kanamori L, Xu C, Hasan SS, Doi SA. Quality versus risk-of-bias assessment in clinical research. *J Clin Epidemiol*. 2021;129:172–175.
2. Hartling L, Hamm M, Milne A, et al. *Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments*. AHRQ Methods for Effective Health Care; 2012.
3. Stone JC, Glass K, Clark J, Ritskes-Hoitinga M, Munn Z, Tugwell P, et al. The MethodologicAI Standards for Epidemiological Research (MASTER) scale demonstrated a unified framework for bias assessment. *J Clin Epidemiol*. 2021;134:52–64.
4. Stone JC, Glass K, Munn Z, Tugwell P, Doi SAR. Comparison of bias adjustment methods in meta-analysis suggests that quality effects modeling may have less limitations than other approaches. *J Clin Epidemiol*. 2020;117:36–45.
5. Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? *Physiotherapy*. 2000;86(2), 94–99.
6. Baumgartner TA, Jackson AS. *Measurement for Evaluation in Physical Education and Exercise Science*. Wm. C. Brown Publishers; 1987.
7. Stone JC, Gurunathan U, Aromataris E, et al. Bias assessment in outcomes research: the role of relative versus absolute approaches. *Value Health*. 2021;24(8), 1145–1149.
8. Aboelsoud MM, Siddique O, Morales A, Seol Y, MO Al-Qadi. fluid choice matters in critically-ill patients with acute pancreatitis: lactated Ringer's vs. isotonic saline. *R I Med J (2013)*. 2016;99(10), 39–42.
9. Choosakul S, Harinwan K, Chirapongsathorn S, et al. Comparison of normal saline versus Lactated Ringer's solution for fluid resuscitation in patients with mild acute pancreatitis, A randomized controlled trial. *Pancreatol*. 2018;18(5), 507–512.
10. de-Madaria E, Herrera-Marante I, Gonzalez-Camacho V, et al. Fluid resuscitation with lactated Ringer's solution vs normal saline in acute pancreatitis: a triple-blind, randomized, controlled trial. *United European Gastroenterol J*. 2018;6(1), 63–72.
11. Karki B, Thapa S, Khadka D, et al. Intravenous Ringers lactate versus normal saline for predominantly mild acute pancreatitis in a Nepalese Tertiary Hospital. *PLoS One*. 2022;17(1), e0263221.
12. Kayhan S, Selcan Akyol B, Ergul M, Baysan C. The effect of type of fluid on disease severity in acute pancreatitis treatment. *Eur Rev Med Pharmacol Sci*. 2021;25(23), 7460–7467.
13. Lee A, Ko C, Buitrago C, et al. Lactated Ringers vs normal saline resuscitation for mild acute pancreatitis: a randomized trial. *Gastroenterology*. 2021;160(3), 955–957. e4.
14. Lipinski M, Rydzewska-Rosolowska A, Rydzewski A, Rydzewska G. Fluid resuscitation in acute pancreatitis: normal saline or lactated Ringer's solution? *World J Gastroenterol*. 2015;21(31), 9367–9372.
15. Olson K, Twohig P, Sayles H, Eichele D, Antoniak D. S3209 Choice of initial resuscitation fluid in patients admitted for acute pancreatitis. *Offic J Am Coll Gastroenterol | ACG*. 2021;116:S1320–S1320.
16. Reddy YR, Talukder S, Yadav TD, Siddappa PK, Kochhar R. Effect of intravenous fluid resuscitation on inflammatory markers of acute pancreatitis and its clinical outcome. *United European Gastroenterol J*. 2014;2(1), 132–135.
17. Vasu De Van P, Verma GR, Bhalla A, et al. Does the type of fluid used in resuscitation matter in the clinical course of acute pancreatitis? *Indian J Gastroenterol*. 2013;32(1), A110.
18. Wu BU, Hwang JQ, Gardner TH, et al. Lactated Ringer's solution reduces systemic inflammation compared with saline in patients with acute pancreatitis. *Clin Gastroenterol Hepatol*. 2011;9(8), 710–717. e1.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.