SYSTEMATIC LITERATURE REVIEW EVALUATING QUALITY ASSESSMENT TOOLS

Cadarette SM, Douyon L, Ranganathan P, Ballew NG, Colby JA, Maiese BA, Slaff S, Wissinger E, Ruiz K

Xcenda, L.L.C., Palm Harbor, FL, USA

Background

- Quality assessment (QA) is an important part of a well-designed systematic literature review (SLR), as assessment of the methodological quality of a study is crucial to ensuring that results are valid (ie, the design and methods have successfully prevented bias).
- Quality assessment tools (QATs) are used to determine the risk of bias in published studies. A wide variety of QATs have been developed to evaluate the design, conduct, and reporting of various types of studies, ranging from cohort studies to randomized controlled trials (RCTs), economic evaluations, and SLRs.
- Finding an appropriate, valid, and easy-to-use QAT for each study design of interest can be a challenge. There is a lack of consensus among researchers as to which tools are best suited for different studies.

Objectives

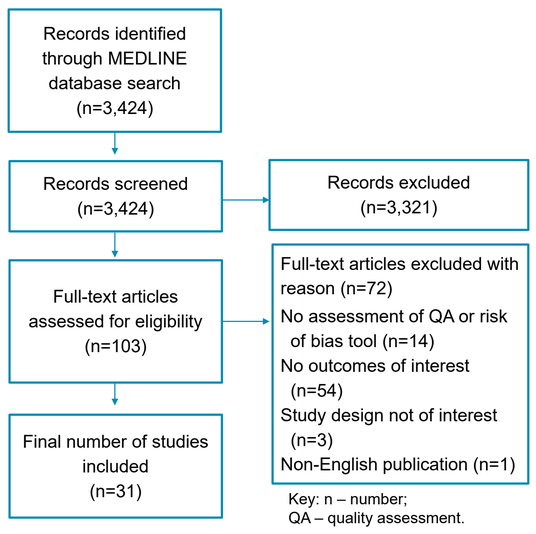
There are 2 objectives of this research:

- To identify published literature evaluating the validity and reliability of QATs
- To assess the validity of currently available QATs for each study design

Methods

- We conducted an SLR to identify published literature evaluating the agreement, validity, and/or reliability within and across QATs.
- A MEDLINE (via PubMed) search was performed from database inception to October 2017 for English-language publications evaluating QATs. A single reviewer performed title/abstract and full-text screening, and a second reviewer conducted the data extraction.
- A PRISMA diagram outlining the literature review process, including the number of records identified, the screening results, and the final number of included references, is provided in **Figure 1**.
- Inclusion and exclusion criteria applied during title/abstract and full-text screening are outlined in **Table 1**.
- QAT validity was assessed via measures of reliability and convergent validity.

Figure 1. Literature Selection and Review Process



Key: n – number; QA – quality assessment.

Table 1: Inclusion and Exclusion Criteria

PICOS	Inclusion	Exclusion
Population	Studies of: ▪ QA tools ▪ Risk of bias assessment and evaluation on human populations in any disease area	▪ No assessment of QA or risk of bias tool ▪ Tool used in animal population
Intervention/Comparator	Not a criterion	None
Outcome	▪ Validity ▪ Reliability ▪ Reproducibility ▪ Agreement	Study does not report at least 1 outcome of interest
Study Design	▪ Primary studies ▪ Systematic review articles	▪ Case reports, editorials, comments, letters, narrative reviews ▪ Non-English publications

Key: PICOS – population, intervention, comparator, outcome, study design; QA – quality assessment.

Results

- Thirty-one studies reporting on 34 scales were included in the SLR, as outlined in **Table 2**.
- The identified QATs most frequently assessed RCTs (25 scales; 15 publications). Five scales (from 4 publications) assessed non-randomized interventional studies, and 3 scales (from 4 publications) assessed SLRs.
- Tools assessing case-control and cohort studies (1 scale; 2 publications), guidelines (1 scale; 1 publication), diagnostic studies (2 scales; 2 publications), health economic studies (1 scale; 1 publication), and meta-analyses (MAs) (1 scale; 1 publication) were also identified.

Table 2: Results by Scales/Study Design

Study Design	Scales Identified			
RCTs	▪ Andrew Scale ▪ Arrive Scale ▪ Balas Scale ▪ Bizzini Scale ▪ CBN RoB ▪ Chalmers Scale ▪ Cho and Bero Scale	▪ Cochrane Collaboration Risk of Bias Tool ▪ CONSORT Scale ▪ Detsky Scale ▪ Downs and Black Scale ▪ EPHPP ▪ GRADE Tool	▪ Imperale Scale ▪ Jadad Scale ▪ MAL ▪ Nguyen Scale ▪ Oxford Pain Validity Scale ▪ PEDro	▪ Reisch Scale ▪ SAQAT ▪ Sindhu Scale ▪ USPSTF System ▪ van Tulder Scale ▪ Yates Scale
Non-randomized interventional studies	▪ Downs and Black Scale ▪ IPM-QRBNR	▪ MINORS	▪ QAREL	▪ USPSTF System
SLRs	▪ AMSTAR	▪ PEDro	▪ QUADAS	
Case-control and cohort studies	▪ Newcastle Ottawa Scale			
Guidelines	▪ MiChe			
MAs	▪ Newcastle Ottawa Scale			
Diagnostic studies	▪ GRADE Tool ▪ QUADAS			
Health economic studies	▪ QHES			

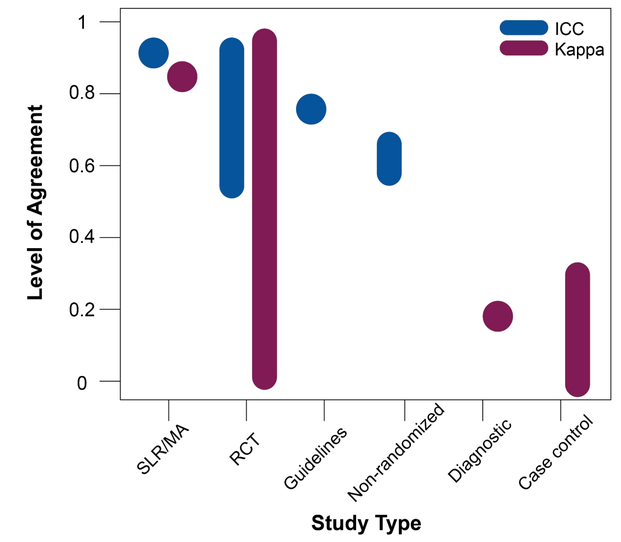
Key: AMSTAR – Assessing the Methodological Quality of Systematic Reviews; CBN RoB – Cochrane Back and Neck Risk of Bias; CONSORT – Consolidated Standards of Reporting Trials; EPHPP – Effective Public Heath Practice Project Quality Assessment Tool; GRADE – Grading of Recommendations Assessment, Development, and Evaluation; IPM-QRBNR – Interventional Pain Management Techniques—Quality Appraisal of Reliability and Risk of Bias Assessment for Non-Randomized Studies; MAs – meta-analyses; MAL – Maastricht-Amsterdam List; MiChe – Mini-Checklist; MINORS – Methodological Index for Non-Randomized Studies; PEDro – Physiotherapy Evidence Database; QAREL – Quality Appraisal of Reliability Studies; QHES – Quality of Health Economic Studies; QUADAS – Quality Assessment of Diagnostic Accuracy Studies; RCTs – randomized controlled trials; SAQAT – semi-automated quality assessment tool; SLRs – systematic literature reviews; USPSTF – United States Preventative Services Task Force.

Results (cont.)

Reliability

- Inter-rater reliability was the most commonly reported measure of agreement within QATs and was measured with the kappa statistic or intraclass correlation coefficient (ICC) (**Figure 2**).
- Inter-rater reliability was high for QATs used with SLRs/MAs (Assessing the Methodological Quality of Systematic Reviews [AMSTAR], Quality Assessment of Diagnostic Accuracy Studies [QUADAS]) (ICC 0.91; kappa 0.84)[1,2] and moderately high for the MiChe scale used with guidelines (ICC 0.76)[3] and for the QATs used with non-randomized interventional studies (Interventional Pain Management Techniques—Quality Appraisal of Reliability and Risk of Bias Assessment for Non-Randomized Studies [IPM-QRBNR], Physiotherapy Evidence Database [PEDro]) (ICC 0.60–0.66).[4,5]
- QATs used with RCTs (**Table 2**) had variable inter-rater reliability scores (ICC 0.55–0.92; kappa 0.02–0.94).[5-11]

Figure 2. Inter-Rater Reliability (Range) of QA Tools Across Study Designs



Key: ICC – intraclass correlation coefficient; MA – meta-analysis; QA – quality assessment; RCT – randomized controlled trial; SLR – systematic literature review.

- QATs for use with diagnostic studies (GRADE) and case-control and cohort studies (Newcastle-Ottawa Scale, GRADE) each had low levels of inter-rater reliability (kappa values of 0.18[12] and 0.00–0.29,[13,14] respectively).

Convergent Validity

- Agreement between tools was reported in 7 publications and was predominantly assessed for QATs used for RCTs (PEDro, Cochrane Risk of Bias, Cochrane Back and Neck [CBN] Risk of Bias [RoB], Effective Public Heath Practice Project Quality Assessment Tool [EPHPP], Van Tulder, Jadad, Downs and Black, United States Preventative Services Task Force [USPSTF]), and for QATs used with SLRs (AMSTAR, R-AMSTAR, Global Assessment Instrument; **Table 3**).
- The highest agreement was seen between variations of the same scale (AMSTAR and R-AMSTAR), which reported kappa 0.89; 95% confidence interval (CI): 0.77, 0.95 (P<0.0001).[15]
- Across other scales assessing RCTs, agreement ranged from poor to fair:
  - The Cochrane RoB and the EPHPP showed very low agreement (kappa 0.006),[16] while the PEDro and the Jadad showed poor correlation as well (ICC 0.35; 95% CI 0.16, 0.54).[17]
  - Moderate agreement was seen between the PEDro and the CBN RoB (kappa 0.61; 95% CI: 0.46, 0.72), especially when adjusted for reliability (kappa 0.83; 95% CI: 0.76, 0.88).[11]
  - One study that evaluated agreement by score level on the PEDro showed evidence that agreement may be stronger for higher-quality trials (kappa range was 0.12–0.44 for agreement with Cochrane RoB for PEDro scores ≥5 to ≥8).[18]

Table 3: Agreement Between Quality Assessment Tools

Author Year	Tool 1	Tool 2	Point Estimate (95% CI)
RCTs			
Armijo-Olivo 2015[18]	PEDro scores ≥5	Cochrane Risk of Bias	Kappa: 0.12 (0.07, 0.16)
Armijo-Olivo 2015[18]	PEDro scores ≥6	Cochrane Risk of Bias	Kappa: 0.24 (0.16, 0.32)
Armijo-Olivo 2015[18]	PEDro scores ≥7	Cochrane Risk of Bias	Kappa: 0.39 (0.286, 0.51)
Armijo-Olivo 2015[18]	PEDro scores ≥8	Cochrane Risk of Bias	Kappa: 0.44 (0.314, 0.574)
Armijo-Olivo 2012[6]	Cochrane Risk of Bias	EPHPP	Kappa: 0.006 (NR)
Yamato 2017[11]	PEDro	CBN RoB	Kappa: 0.61 (0.46, 0.72)
Yamato 2017[11]	PEDro	CBN RoB	Kappa*: 0.83 (0.76, 0.88)
Macedo 2010[17]	PEDro	Van Tulder 2003	ICC: 0.71 (0.41, 0.95)
Macedo 2010[17]	PEDro	Jadad	ICC: 0.35 (0.16, 0.54)
RCTs/Non-Randomized Interventional Studies			
O'Connor 2015[19]	Downs and Black	USPSTF	ICC: 0.8 (0.57, 0.92)
SLRs			
Popovich 2012[15]	R-AMSTAR	AMSTAR	R_S: 0.89 (0.77, 0.95)
Popovich 2012[15]	R-AMSTAR	AMSTAR	R_S: 0.53 (0.21, 0.75)
Shea 2007[2]	AMSTAR	Global Assessment Instrument	R: 0.72 (0.53, 0.84)

Key: AMSTAR – Assessing the Methodological Quality of Systematic Reviews; CBN RoB – Cochrane Back and Neck Risk of Bias; CI – confidence interval; EPHPP – Effective Public Heath Practice Project Quality Assessment Tool; ICC – intraclass correlation coefficient; NR – not reported; PEDro – Physiotherapy Evidence Database; R – Pearson's correlation coefficient; R_s – Spearman's correlation coefficient; R-AMSTAR – Revised Assessing the Methodological Quality of Systematic Reviews; RCT – randomized controlled trial; SLR – systematic literature review; USPSTF – United States Preventative Services Task Force.
*Adjusted for reliability.

Conclusions

- Best practices call for the use of validated QATs to assess all of the literature included in SLRs, which frequently extend beyond RCTs to encompass an array of study designs.
- The inter-rater reliability and agreement results with currently available scales are highly dependent on both the choice of scale and the individual scoring it.
- We found published evidence on the validity of QATs to be limited, especially for study designs other than RCTs, making it difficult to identify validated tools for use in SLRs.
- Further, the QATs included in our study predominantly demonstrated low to moderate levels of validity.
- Although high inter-rater reliability was reported for some RCT, SLR, and guideline QATs, each of the high-scoring QATs was limited to 1 study per QAT, limiting the wider applicability of this evidence.
  - Additionally, reliability and convergent validity were variable for RCT QATs and fair to poor for other study types.
- Bridging the gap between current practices and best practices for identifying potential sources of bias will require: 1) training raters to more consistently apply QATs; and 2) developing and validating new QATs.

References

1. Henschke N, et al. *J Clin Epidemiol*. 2014 Apr;67(4):416-424.

2. Shea BJ, et al. *PLoS One*. 2007;2(12):e1350.

3. Siebenhofer A, et al. *BMC Med Res Methodol*. 2016;16:39.

4. Manchikanti L, et al. *Pain Physician*. 2014 May-Jun;17(3): E291-317.

5. Murray E, et al. *Int J Lang Commun Disord*. 2013 May-Jun;48(3):297-306.

6. Armijo-Olivo S, et al. *J Eval Clin Pract*. 2012 Feb;18(1):12-18.

7. Hartling L, et al. *PLoS One*. 2011;6(2):e17242.

8. Hartling L, et al. *J Clin Epidemiol*. 2013a Sep;66(9):973-981.

9. Maher CG, et al. *Phys Ther*. 2003 Aug;83(8):713-721.

10. Olivo SA, et al. *Phys Ther*. 2008 Feb;88(2):156-175.

11. Yamato TP, et al. *J Clin Epidemiol*. 2017 Jun;86:176-181.

12. Hollingworth W, et al. *Acad Radiol*. 2006 Jul;13(7):803-810.

13. Hartling L, et al. *J Clin Epidemiol*. 2013b Sep;66(9):982-993.

14. Lo CK, et al. *BMC Med Res Methodol*. 2014;14:45.

15. Popovich I, et al. *PLoS One*. 2012;7(12):e50403.

16. Hartling L, et al. *PLoS One*. 2012;7(4):e34697.

17. Macedo LG, et al. *J Clin Epidemiol*. 2010 Aug;63(8):920-925.

18. Armijo-Olivo S, et al. *PLoS One*. 2015;10(7):e0132634.

19. O'Connor SR, et al. *BMC Res Notes*. 2015;8:224.

Presented at ISPOR 21st Annual European Congress | November 14, 2018 | Barcelona, Spain | Funded by Xcenda

Xcenda® AmerisourceBergen

SYSTEMATIC LITERATURE REVIEW (SLR) EVALUATING QUALITY ASSESSMENT TOOLS (QAT)

Cadarette SM, Douyon L, Ranganathan P, Ballew NG, Colby JA, Maiese BA, Slaff S, Wissinger E, Ruiz K

Xcenda, L.L.C., Palm Harbor, FL, USA

Background

- Quality assessment (QA) is an important part of a well-designed systematic literature review (SLR), as assessment of the methodological quality of a study is crucial to ensuring that results are valid (ie, the design and methods have successfully prevented bias).
- Quality assessment tools (QATs) are used to determine the risk of bias in published studies. A wide variety of QATs have been developed to evaluate the design, conduct, and reporting of various types of studies, ranging from cohort studies to randomized controlled trials (RCTs), economic evaluations, and SLRs.
- Finding an appropriate, valid, and easy-to-use QAT for each study design of interest can be a challenge. There is a lack of consensus among researchers as to which tools are best suited for different studies.

Objectives

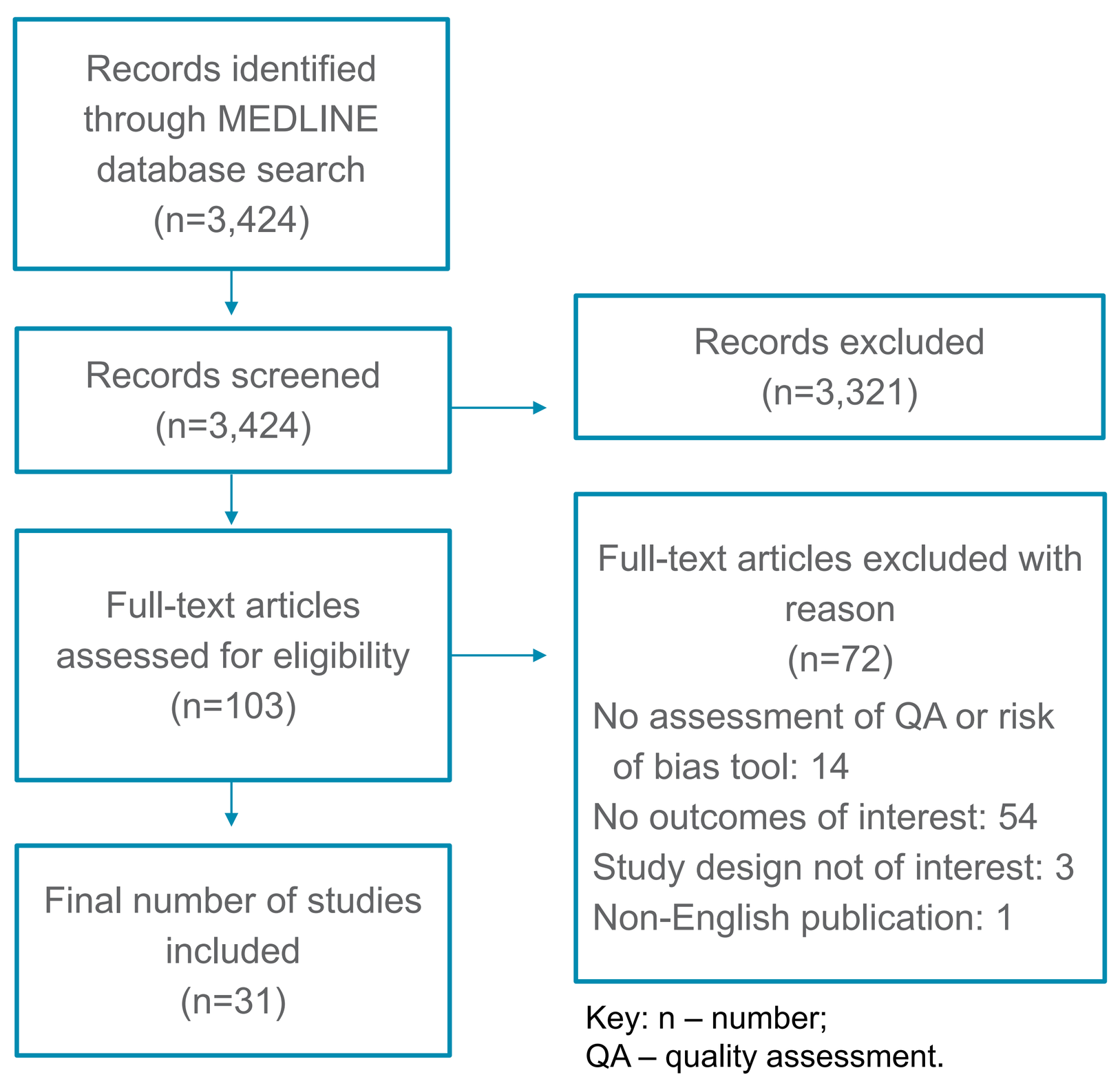
There are 2 objectives of this research:

- To identify published literature evaluating the validity and reliability of QATs
- To assess the validity of currently available QATs for each study design

Methods

- We conducted an SLR to identify published literature evaluating the agreement, validity, and/or reliability within and across QATs.
- A MEDLINE (via PubMed) search was performed from database inception to October 2017 for English language publications evaluating QATs. A single reviewer performed title/abstract and full-text screening, and a second reviewer conducted the data extraction.
- A PRISMA diagram outlining the literature review process, including the number of records identified, the screening results, and the final number of included references, is provided in **Figure 1**.
- Inclusion and exclusion criteria applied during title/abstract and full-text screening are outlined in **Table 1**.
- QAT validity was assessed via measures of reliability and convergent validity.

Figure 1. Literature Selection and Review Process



Key: n – number; QA – quality assessment.

Table 1: Inclusion and Exclusion Criteria

PICOS	Inclusion	Exclusion
Population	Studies of: ▪ QA tools ▪ Risk of bias assessment and evaluation on human populations in any disease area	▪ No assessment of QA or risk of bias tool ▪ Tool used in animal population
Intervention/Comparator	Not a criterion	None
Outcome	▪ Validity ▪ Reliability ▪ Reproducibility ▪ Agreement	Study does not report at least 1 outcome of interest
Study Design	▪ Primary studies ▪ Systematic review articles	▪ Case reports, editorials, comments, letters, narrative reviews ▪ Non-English publications

Key: PICOS – population, intervention, comparator, outcome, study design; QA – quality assessment.

Results

- Thirty-one studies reporting on 34 scales were included in the SLR, as outlined in **Table 2**.
- The identified QATs most frequently assessed RCTs (25 scales; 15 publications). Five scales (from 4 publications) assessed non-randomized interventional studies, and 3 scales (from 4 publications) assessed SLRs.
- Tools assessing case-control and cohort studies (1 scale; 2 publications), guidelines (1 scale; 1 publication), diagnostic studies (2 scales; 2 publications), health economic studies (1 scale; 1 publication), and meta-analyses (MAs) (1 scale; 1 publication) were also identified.

Table 2: Results by Scales/Study Design

Study Design	Scales Identified
RCTs	▪ Andrew Scale ▪ Arrive Scale ▪ Balas Scale ▪ Bizzini Scale ▪ CBN RoB ▪ Chalmers Scale ▪ Cho and Bero Scale ▪ Cochrane Collaboration Risk of Bias Tool ▪ CONSORT Scale ▪ Detsky Scale ▪ Downs and Black Scale ▪ EPHPP ▪ GRADE Tool ▪ Imperale Scale ▪ Jadad Scale ▪ MAL ▪ Nguyen Scale ▪ Oxford Pain Validity Scale ▪ PEDro ▪ Reisch Scale ▪ SAQAT ▪ Sindhu Scale ▪ USPSTF System ▪ van Tulder Scale ▪ Yates Scale
Non-randomized interventional studies	▪ Downs and Black Scale ▪ IPM-QRBNR ▪ MINORS ▪ QAREL ▪ USPSTF System
SLRs	▪ AMSTAR ▪ PEDro ▪ QUADAS
Case-control and cohort studies	Newcastle Ottawa Scale
Guidelines	MiChe
MAs	Newcastle Ottawa Scale
Diagnostic studies	▪ GRADE Tool ▪ QUADAS
Health economic studies	QHES

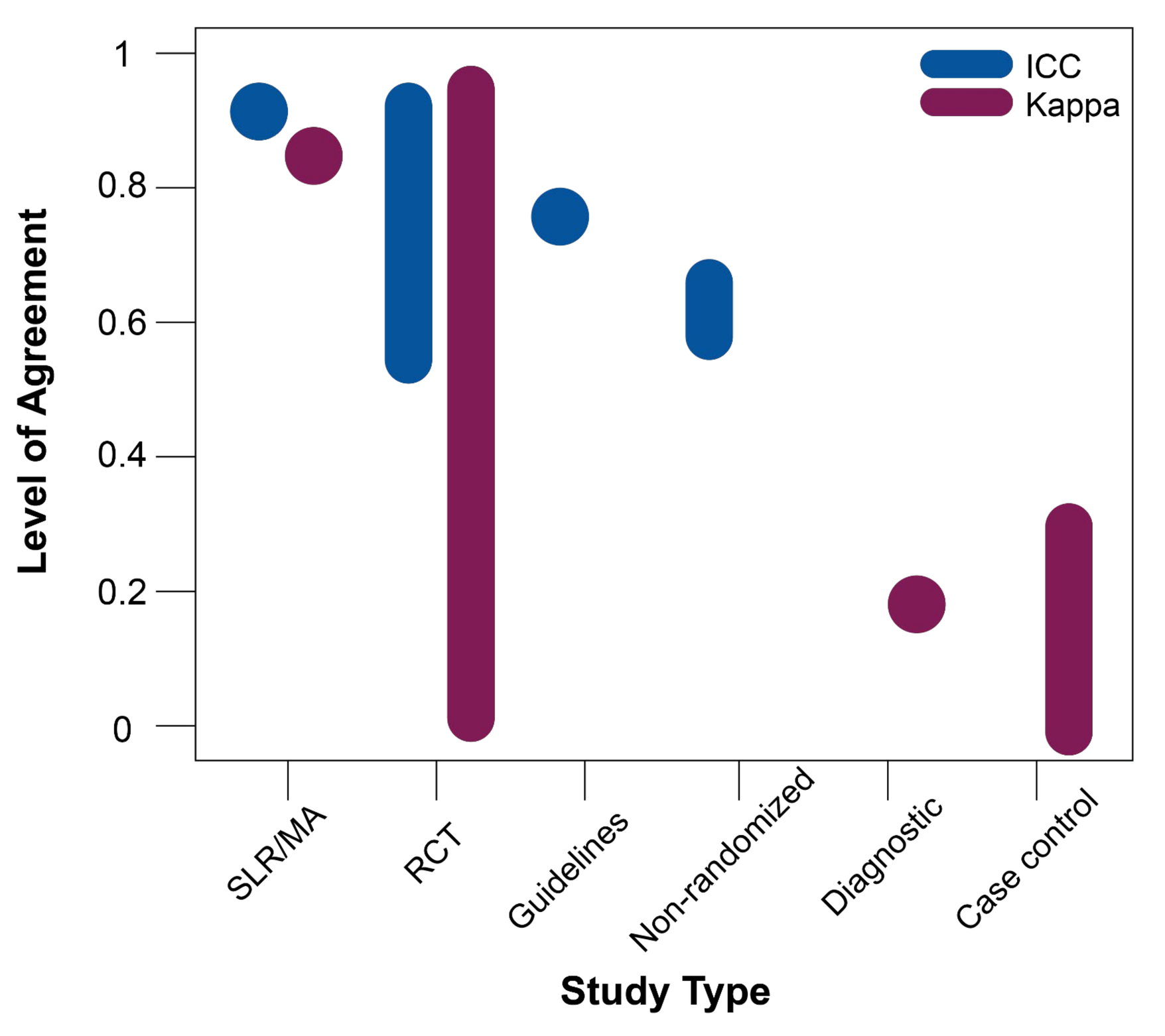
Key: AMSTAR – Assessing the Methodological Quality of Systematic Reviews; CBN RoB – Cochrane Back and Neck Risk of Bias; CONSORT – Consolidated Standards of Reporting Trials; EPHPP – Effective Public Heath Practice Project Quality Assessment Tool; GRADE – Grading of Recommendations Assessment, Development, and Evaluation; IPM-QRBNR – Interventional Pain Management Techniques—Quality Appraisal of Reliability and Risk of Bias Assessment for Non-Randomized Studies; MAs – meta-analyses; MAL – Maastricht-Amsterdam List; MiChe – Mini-Checklist; MINORS – Methodological Index for Non-Randomized Studies; PEDro – Physiotherapy Evidence Database; QAREL – Quality Appraisal of Reliability Studies; QHES – Quality of Health Economic Studies; QUADAS – Quality Assessment of Diagnostic Accuracy Studies; RCTs – randomized controlled trials; SAQAT – semi-automated quality assessment tool; SLRs – systematic literature reviews; USPSTF – United States Preventative Services Task Force.

Reliability

- Inter-rater reliability was the most commonly reported measure of agreement within QATs and was measured with the kappa statistic or intraclass correlation coefficient (ICC) (**Figure 2**).
- Inter-rater reliability was high for QATs used with SLRs/MAs (Assessing the Methodological Quality of Systematic Reviews [AMSTAR], Quality Assessment of Diagnostic Accuracy Studies [QUADAS]) (ICC 0.91; kappa 0.84)[1,2] and moderately high for the MiChe scale used with guidelines (ICC 0.76)[3] and for the QATs used with non-randomized interventional studies (Interventional Pain Management Techniques—Quality Appraisal of Reliability and Risk of Bias Assessment for Non-Randomized Studies [IPM-QRBNR], Physiotherapy Evidence Database [PEDro]) (ICC 0.60–0.66).[4,5]
- QATs used with RCTs (**Table 2**) had variable inter-rater reliability scores (ICC 0.55–0.92; kappa 0.02–0.94).[5-11]

Results (cont.)

Figure 2. Inter-Rater Reliability (Range) of QA Tools Across Study Designs



- QATs for use with diagnostic studies (GRADE) and case-control and cohort studies (Newcastle-Ottawa Scale, GRADE) each had low levels of inter-rater reliability (kappa values of 0.18[12] and 0.00–0.29,[13,14] respectively).

Convergent Validity

- Agreement between tools was reported in 7 publications and was predominantly assessed for QATs used for RCTs (PEDro, Cochrane Risk of Bias, Cochrane Back and Neck [CBN] Risk of Bias [RoB], Effective Public Heath Practice Project Quality Assessment Tool [EPHPP], Van Tulder, Jadad, Downs and Black, United States Preventative Services Task Force [USPSTF]), and for QATs used with SLRs (AMSTAR, R-AMSTAR, Global Assessment Instrument; **Table 3**).
- The highest agreement was seen between variations of the same scale (AMSTAR and R-AMSTAR), which reported kappa 0.89; 95% confidence interval (CI): 0.77, 0.95 (P<0.0001).[15]
- Across other scales assessing RCTs, agreement ranged from poor to fair:
  - The Cochrane RoB and the EPHPP showed very low agreement (kappa 0.006),[16] while the PEDro and the Jadad showed poor correlation as well (ICC 0.35; 95% CI 0.16, 0.54).[17]
  - Moderate agreement was seen between the PEDro and the CBN RoB (kappa 0.61; 95% CI: 0.46, 0.72), especially when adjusted for reliability (kappa 0.83; 95% CI: 0.76, 0.88).[11]
  - One study that evaluated agreement by score level on the PEDro showed evidence that agreement may be stronger for higher-quality trials (kappa range was 0.12–0.44 for agreement with Cochrane RoB for PEDro scores ≥5 to ≥8).[18]

Table 3: Agreement Between Quality Assessment Tools

Author Year	Tool 1	Tool 2	Statistic	95% CI (P)
RCTs				
ARMIJO-OLIVO 2015[18]	PEDRO SCORES ≥5	COCHRANE RISK OF BIAS	KAPPA: 0.12	0.07, 0.16
ARMIJO-OLIVO 2015[18]	PEDRO SCORES ≥6	Cochrane Risk of Bias	Kappa: 0.24	0.16, 0.32
ARMIJO-OLIVO 2015[18]	PEDRO SCORES ≥7	Cochrane Risk of Bias	Kappa: 0.39	0.286, 0.51
ARMIJO-OLIVO 2015[18]	PEDRO SCORES ≥8	Cochrane Risk of Bias	Kappa: 0.44	0.314, 0.574
ARMIJO-OLIVO 2012[6]	COCHRANE RISK OF BIAS	EPHPP	Kappa: 0.006	NR
YAMATO 2017[11]	PEDRO	CBN RoB	Kappa: 0.61	0.46, 0.72
YAMATO 2017[11]	PEDRO	CBN RoB	Kappa*: 0.83	0.76, 0.88
MACEDO 2010[17]	PEDRO	Van Tulder 2003	ICC: 0.71	0.41, 0.95
MACEDO 2010[17]	PEDRO	Jadad	ICC: 0.35	0.16, 0.54
RCTs/Non-Randomized Interventional Studies				
O'CONNOR 2015[19]	Downs and black	USPSTF	Icc: 0.8	0.57, 0.92
SLRs				
POPOVICH 2012[15]	R-AMSTAR	AMSTAR	R_S: 0.89	0.77, 0.95 (P<0.0001)
POPOVICH 2012[15]	R-AMSTAR	AMSTAR	R_S: 0.53	0.21, 0.75 (P=0.0029)
SHEA 2007[2]	AMSTAR	Global Assessment Instrument	r: 0.72	0.53, 0.84

Key: AMSTAR – Assessing the Methodological Quality of Systematic Reviews; CBN RoB – Cochrane Back and Neck Risk of Bias; CI – confidence interval; EPHPP – Effective Public Heath Practice Project Quality Assessment Tool; ICC – intraclass correlation coefficient; NR – not reported; PEDro – Physiotherapy Evidence Database; R-AMSTAR – Revised Assessing the Methodological Quality of Systematic Reviews; RCT – randomized controlled trial; SLR – systematic literature review; USPSTF – United States Preventative Services Task Force.
*Adjusted for reliability.
**Construct validity.

Conclusions

- Best practices call for the use of validated QATs to assess all of the literature included in SLRs, which frequently extend beyond RCTs to encompass an array of study designs.
- The inter-rater reliability and agreement results with currently available scales are highly dependent on both the choice of scale and the individual scoring it.
- We found published evidence on the validity of QATs to be limited, especially for study designs other than RCTs, making it difficult to identify validated tools for use in SLRs.
- Further, the QATs included in our study predominantly demonstrated low to moderate levels of validity.
- Although high inter-rater reliability was reported for some RCT, SLR, and guidelines QATs, each of the high-scoring QATs was limited to 1 study per QAT, limiting the wider applicability of this evidence.
  - Additionally, reliability and convergent validity were variable for RCT QATs and fair to poor for other study types.
- Bridging the gap between current practices and best practices for identifying potential sources of bias will require: 1) training raters to more consistently apply QATs; and 2) developing and validating new QATs.

1. Hensonke N, Keuerleber J, Ferreira M, Maher CG, Verhagen AP. The methodological quality of diagnostic test accuracy studies for musculoskeletal conditions can be improved. *J Clin Epidemiol*. 2014;67(4):416-424.
2. Shea BJ, Bouter LM, Peterson J, et al. External validation of a measurement tool to assess systematic reviews (AMSTAR). *PLoS One*. 2007;2(12):e1350.
3. Siebenhofer A, Semlitsch T, Herborn T, Siering U, Kopp I, Hartig J. Validation and reliability of a guideline appraisal mini-checklist for daily practice use. *BMC Med Res Methodol*. 2016;16:39.
4. Manchikanti L, Hirsch JA, Heavner JE, et al. Development of an interventional pain management specific instrument for methodologic quality assessment of nonrandomized studies of interventional techniques. *Pain Physician*. 2014;17(3):E291-317.
5. Murray E, Power E, Togher L, McCabe P, Munro N, Smith K. The reliability of methodological ratings for speechBITE using the PEDro-P scale. *Int J Lang Commun Disord*. 2013;48(3):297-306.
6. Armijo-Olivo S, Stiles CR, Hagen NA, Biondo PD, Cummings GG. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. *J Eval Clin Pract*. 2012;18(1):12-18.
7. Hartling L, Bond K, Vandermeer B, Seida J, Dryden DM, Rowe BH. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. *PLoS One*. 2011;6(2):e17242.
8. Hartling L, Hamm MP, Milne A, 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(9):973-981.
9. Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M. Reliability of the PEDro scale for rating quality of randomized controlled trials. *Phys Ther*. 2003;83(8):713-721.
10. Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. Scales to assess the quality of randomized controlled trials: a systematic review. *Phys Ther*. 2008;88(2):156-175.
11. Yamato TP, Maher C, Koes B, Moseley A. The PEDro scale had acceptably high convergent validity, construct validity, and interrater reliability in evaluating methodological quality of pharmaceutical trials. *J Clin Epidemiol*. 2017;86:176-181.
12. Hollingworth W, Medina LS, Lenkinski RE, et al. Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool. A preliminary assessment. *Acad Radiol*. 2006;13(7):803-810.
13. Hartling L, Milne A, Hamm MP, et al. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. *J Clin Epidemiol*. 2013;66(9):982-993.
14. Lo CK, Mertz D, Loeb M. Newcastle-Ottawa Scale: comparing reviewers' to authors' assessments. *BMC Med Res Methodol*. 2014;14:45.
15. Popovich I, Windsor B, Jordan V, Showell M, Shea B, Farquhar CM. Methodological quality of systematic reviews in subfertility: a comparison of two different approaches. *PLoS One*. 2012;7(12):e50403.
16. Hartling L, Fernandes RM, Seida J, Vandermeer B, Dryden DM. From the trenches: a cross-sectional study applying the GRADE tool in systematic reviews of healthcare interventions. *PLoS One*. 2012;7(4):e34697.
17. Macedo LG, Elkins MR, Maher CG, Moseley AM, Herbert RD, Sherrington C. There was evidence of convergent and construct validity of Physiotherapy Evidence Database quality scale for physiotherapy trials. *J Clin Epidemiol*. 2010;63(8):920-925.
18. Armijo-Olivo S, da Costa BR, Cummings GG, et al. PEDro or Cochrane to assess the quality of clinical trials? A meta-epidemiological study. *PLoS One*. 2015;10(7):e0132634.
19. O'Connor SR, Tully MA, Ryan B, Bradley JM, Baxter GD, McDonough SM. Failure of a numerical quality assessment scale to identify potential risk of bias in a systematic review: a comparison study. *BMC Res Notes*. 2015;8:224.

Presented at ISPOR 21st Annual European Congress | November 14, 2018 | Barcelona, Spain | Funded by Xcenda

Xcenda AmerisourceBergen