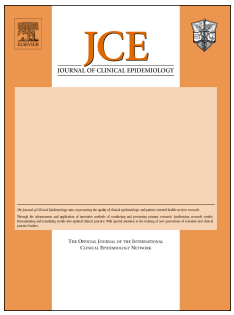

Common challenges and suggestions for risk of bias tool development: a systematic review of methodological studies

Eve Tomlinson, Chris Cooper, Clare Davenport, Anne W.S. Rutjes, Mariska Leeflang, Sue Mallett, Penny Whiting

PII: S0895-4356(24)00125-2

DOI: https://doi.org/10.1016/j.jclinepi.2024.111370

Reference: JCE 111370

To appear in: *Journal of Clinical Epidemiology*

Received Date: 23 November 2023

Revised Date: 16 April 2024

Accepted Date: 17 April 2024

Please cite this article as: Tomlinson E, Cooper C, Davenport C, Rutjes AWS, Leeflang M, Mallett S, Whiting P, Common challenges and suggestions for risk of bias tool development: a systematic review of methodological studies, *Journal of Clinical Epidemiology* (2024), doi: https://doi.org/10.1016/j.jclinepi.2024.111370.

Common challenges and suggestions for risk of bias tool development: a systematic review of methodological studies

Eve Tomlinson[a], Chris Cooper[a], Clare Davenport[b,c], Anne W.S. Rutjes[d], Mariska Leeflang[e], Sue Mallett[f], Penny Whiting[a]

[a] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK.
[b] Test and Prediction Group, Institute of Applied Health Research, University of Birmingham, Birmingham, B15 2TT, UK
[c] NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham B15 2TT, UK
[d] Department of Medical and Surgical Sciences for Children and Adults (SMECHIMAI), University of Modena and Reggio Emilia, Modena, Italy
[e] Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands
[f] Centre for Medical Imaging, University College London, London, UK

Corresponding author:
Eve Tomlinson
Population Health Sciences,
Bristol Medical School,
University of Bristol,
Canynge Hall, 39 Whatley Road,
Bristol, BS8 2PS,
United Kingdom.
Email: eve.tomlinson@bristol.ac.uk

Abstract

Objective: To review the findings of studies that have evaluated the design and/or usability of key risk of bias (RoB) tools for the assessment of RoB in primary studies, as categorised by the LATITUDES Network (a searchable library of RoB tools for evidence synthesis): PROBAST, RoB2, ROBINS-I, QUADAS-2, QUADAS-C, QUAPAS, ROBINS-E, and the COSMIN RoB checklist.

Study design and setting: Systematic review of methodological studies. We conducted a forward citation search from the primary report of each tool, to identify primary studies aiming to evaluate the design and/or usability of the tool. Two reviewers assessed studies for inclusion. We extracted tool features into Microsoft Word and used NVivo for document analysis, comprising a mix of deductive and inductive approaches. We summarised findings within each tool and explored common findings across tools.

Results: We identified 13 tool evaluations meeting our inclusion criteria: PROBAST (3); RoB2 (3); ROBINS-I (4); QUADAS-2 (3). We identified no evaluations for the other tools. Evaluations varied in clinical topic area, methodology, approach to bias assessment and tool user background. Some had limitations affecting generalisability. We identified common findings across tools for 6/14 themes: 1) challenging items (e.g. RoB2/ROBINS-I "deviations from intended interventions" domain), 2) overall RoB judgement (concerns with overall risk calculation in PROBAST/ROBINS-I), 3) tool usability (concerns about complexity), 4) time to complete tool (varying demands on time e.g. depending on number of outcomes assessed), 5) user agreement (varied across tools), and 6) recommendations for future use (e.g. piloting) and development (add intermediate domain answer to QUADAS-2/PROBAST; provide clearer guidance for all tools). Of the other eight themes, seven only had findings for the QUADAS-2 tool, limiting comparison across tools, and one ("re-organisation of questions") had no findings.

Conclusion: Evaluations of key RoB tools have posited common challenges and recommendations for tool use and development. These findings may be helpful to people using or developing RoB

tools. Guidance is necessary to support the design and implementation of future RoB tool evaluations.

Running title: Systematic review of methodological studies

Keywords: Risk of bias, Systematic reviews, Evaluation, Research methods, RoB

Manuscript word count: 3246

Plain language summary

What is the problem?

Systematic reviews are a type of research that bring together the results of individual studies to answer a research question. The results of systematic reviews are used by people (e.g. doctors, patients and policy makers) to help them make important decisions about health care.

The reliability of a systematic review depends on how well the studies included in the review have been conducted. If the included studies have not been conducted well, then their results might be "biased" i.e. not show the truth. This is a problem because a systematic review of biased studies may give a misleading conclusion.

Researchers who do systematic reviews are advised to use a "risk of bias tool" to assess whether the included studies are biased. It is not clear whether researchers find risk of bias tools easy to use, or whether they could be improved.

What did we do?

We identified studies that had evaluated the design and/or ease of use of commonly used risk of bias tools. We extracted study findings, including parts of the tools that researchers find hard to use and suggestions for how the tools could be improved.

What did we find?

Thirteen studies had evaluated four risk of bias tools: PROBAST, RoB2, ROBINS-I, and QUADAS-2. We found no studies that had evaluated the following risk of bias tools: QUADAS-C, QUAPAS, ROBINS-E, or the COSMIN RoB checklist. The studies found that the tools can be hard to use and that they could be improved e.g. by making them simpler and by creating clearer guidance. Studies also gave recommendations for people using the tools in the future. These findings will be helpful to people who develop risk of bias tools and people who use the tools.

What is new?

Key findings

- Thirteen studies have evaluated the design and usability of risk of bias (RoB) tools: PROBAST, RoB2, ROBINS-I, QUADAS-2. The studies highlight common challenges with tool implementation and provide recommendations for tool use and development. We identified no primary studies evaluating QUADAS-C, QUAPAS, ROBINS-E, or the COSMIN RoB checklist.
- Some of the existing tool evaluation studies have methodological limitations restricting the generalisability of their findings.

What this adds to what is known

- This is the first review to summarise findings from studies evaluating the design and usability of RoB tools.

What is the implication, what should change now

- Our findings will be helpful to tool users and developers who may benefit from other experiences of applying RoB tools and recommendations for improving the RoB assessment process. Anyone planning to evaluate a RoB tool may also benefit from our review of the current literature on tool evaluation, evaluation methodology used and limitations of existing research.
- Guidance is necessary to ensure appropriate design and implementation of RoB tool evaluation studies in future.

1. Introduction

Systematic reviews seek to answer a specific research question by summarising results from primary studies that meet pre-specified eligibility criteria. Results from primary studies may be biased (i.e. deviate from the truth) due to inappropriate design, conduct, analysis and/ or reporting. A review of biased results is likely to provide a misleading conclusion. This is an issue, as conclusions from systematic reviews are widely used to inform clinical, policy and patient decision-making. Guidance for systematic review conduct therefore emphasises the importance of completing risk of bias (RoB) assessment in systematic reviews,(1) and reporting guidelines (such as PRISMA 2020(2)) recommend that reviewers transparently report RoB assessments. This information contributes to certainty in the overall evidence.(3)

Several tools exist to facilitate the assessment of RoB for varying study designs.(4, 5) The LATITUDES Network provides a searchable library of such tools.(6) LATITUDES identifies RoB tools for inclusion in the library through liaison with experts interested in evidence synthesis or RoB assessment, screening existing depositories of tools, screening reviews that have focused on tools, and a review of guidance from evidence synthesis organisations. LATITUDES listed eight key tools designed to assess the RoB in primary studies at launch in September 2023: QUADAS-2 for diagnostic accuracy studies,(7) RoB2 for randomised controlled trials,(8) ROBINS-I for non-randomised studies of interventions,(9) PROBAST for diagnostic and prognostic prediction model studies,(10) QUADAS-C for comparative diagnostic accuracy studies,(11) QUAPAS for prognostic accuracy studies,(12) ROBINS-E for exposure studies,(13) and the COSMIN RoB checklist for studies on measurement properties of patient-reported outcome measures.(14). These tools fulfil specific criteria posited by the LATITUDES Network(6), they: 1) focus on RoB, or distinction between items that assess RoB and other aspects of study quality, 2) offer a method to reach either a domain specific or overall assessment of RoB 3) have been used in at least one review that none of the tool authors were co-authors on or are an update to a previously recommended LATITUDES key tool, 4) have been developed involving collaborators from different disciplines (e.g. methodologists, statisticians, clinicians), and 5) avoid use of summary numerical quality scores.

Reporting guidelines also exist for many of the study designs previously outlined – for example, TRIPOD(15) for the reporting of prediction models, CONSORT(16) for the reporting of randomised controlled trials, STROBE(17) for observational studies and STARD(18) for diagnostic test accuracy studies. These resources differ to risk of bias tools – reporting guidelines are checklists of information used to guide authors of primary studies in reporting a specific type of research, whereas risk of bias tools are used by systematic reviewers to assess the risk of bias of included primary studies. However, the goal of using reporting guidelines is to improve the quality of the reporting of studies,(19) which in turn supports systematic reviewers in conducting risk of bias assessments.

Studies that have evaluated RoB tools have not previously been summarised in a review. Therefore, this systematic review of methodological studies aims to appraise the findings of studies that have evaluated the design and/or usability the following RoB tools: PROBAST, RoB2, ROBINS-I, QUADAS-2, QUADAS-C, QUAPAS, ROBINS-E, and the COSMIN RoB checklist. A summary of the structure (e.g. number of domains, number of signalling questions) and features (e.g. tool purpose, level of assessment, process for generating an overall risk of bias rating) of these RoB tools is provided in supplementary material section 2. The tools have many similarities but also some differences, for example in level of assessment, domain and signalling question answer options, and software to

support use. In this review, we highlight challenges faced by tool users, report suggestions for tool development, and discuss tool evaluation methodology.

2. Methods

This review was conducted in line with published guidance on systematic review conduct,(1) and is reported according to PRISMA 2020 (see supplementary material section 6), with adaptations made for the methodological nature of this review.(20) The protocol was registered on Open Science Framework.(21)

2.1. Eligibility criteria

Studies were eligible for inclusion if they stated in the title or abstract, that they aimed to evaluate the design (e.g. the number of domains, signalling questions or answer options) and/or usability of the most recent version of any of the following tools: QUADAS-2,(7) RoB2,(8) ROBINS-I,(9) PROBAST,(10) QUADAS-C,(11) QUAPAS,(12) ROBINS-E,(13) or the COSMIN RoB checklist.(14). We included studies with quantitative or qualitative findings and did not restrict based on publication language or clinical topic area.

2.2. Search strategy

To identify evaluation studies, we undertook a forward citation search in Science Citations Index Expanded (SCI-Expanded, Clarivate, 1990-current) using the primary report of each tool as the source.(22) Tools were searched in two phases as the LATITUDES Network website was developed. In July 2023, we searched for QUADAS-2, PROBAST, ROBINS-I and RoB2.(7-10) This included a Google Scholar search to identity eligible evaluations not yet included in SCI-expanded. In October 2023, we searched the remaining tools added to the LATITUDES Network website, namely: QUADAS-C, QUAPAS, and the COSMIN RoB checklist.(11, 12, 14) ROBINS-E (13) had not been formally published but we checked the recommended citation for this tool. We also checked the bibliographies of all included evaluation studies.

2.3. Study selection

Two reviewers (ET, CC) undertook study selection. The reviewers independently screened titles and abstracts of reports identified by the searches, using the web-hosted screening tool Rayyan. Full reports of those considered potentially relevant were obtained and assessed for inclusion. Any potentially relevant reports identified by checking the reference lists of included studies were also assessed for inclusion. Any disagreements were resolved by consensus. No automation tools were used.

2.4. Data Extraction

We extracted the following study characteristics from each evaluation study using NVivo(23): first author, year of publication, country of first author, journal, title, aim, model/ study sampling method, details of models/studies assessed (including clinical topic area, study design, number of studies/ models), number of individuals using tool, user background, user experience with tool, guidance used. One reviewer conducted data extraction (ET), and where there was uncertainty, this was checked by a second reviewer (CD). The team met at regular intervals to discuss findings from data extraction.

2.5. Data Synthesis

We tabulated a summary of the included evaluation studies' characteristics and methodology.

We conducted document analysis of the included evaluation studies. One reviewer (ET) used NVivo to extract data: relevant information was highlighted from study reports and assigned to themes. We began with a deductive approach and categorised information into the coding structure prespecified in our protocol (see Table 1). This was developed by the review team, based on what we sought to find in the literature. When we found additional relevant information that did not fit into these themes, we moved to an inductive approach and added new themes as necessary, following discussion among the review team.

Table 1 Themes used to organise data

General findings	Overall usability
	Time to complete the tool
	Experience of generating overall risk of bias judgements
	User agreement
	Floor and ceiling effects*
	Smallest detectable change*
	Standard error of measurement*
	Internal consistency*
Domain and signalling question specific findings	Items to remove
	Items that are described as being challenging
	Items to reword
	Items to add (not study-specific)
	Suggestions for re-organising signalling questions e.g. into different domains
Future use/ development	Recommendations relating to the use or future development of the tool

*These themes were added using an inductive approach

Following document analysis in NVivo, the information was exported into Microsoft word, providing a list of themes and associated quotes from each study. This information constituted our findings from each included evaluation study. We used this to produce: 1) a narrative summary of findings for each study within each included tool, and 2) a narrative summary of findings across tools, highlighting common findings. We tabulated any quantitative data.

2.6. Differences between protocol and review

For completeness, we included all key tools designed to assess the RoB in primary studies listed on the LATITUDES Network website at launch in September 2023, rather than restrict to those considered by LATITUDES before March 2023, as was noted in our protocol.

3. Results

3.1. Search results

Searches identified 19,512 unique reports. Of these, 13 studies (17 reports) were included in the review (Figure 1). This included 3 studies for PROBAST, 3 for RoB2, 3 for QUADAS-2, and 4 for ROBINS-I. We identified no evaluation studies for QUADAS-C, QUAPAS, ROBINS-E or the COSMIN RoB checklist. References for included studies and studies excluded at full text assessment are presented in supplementary material section 1, stratified by tool.

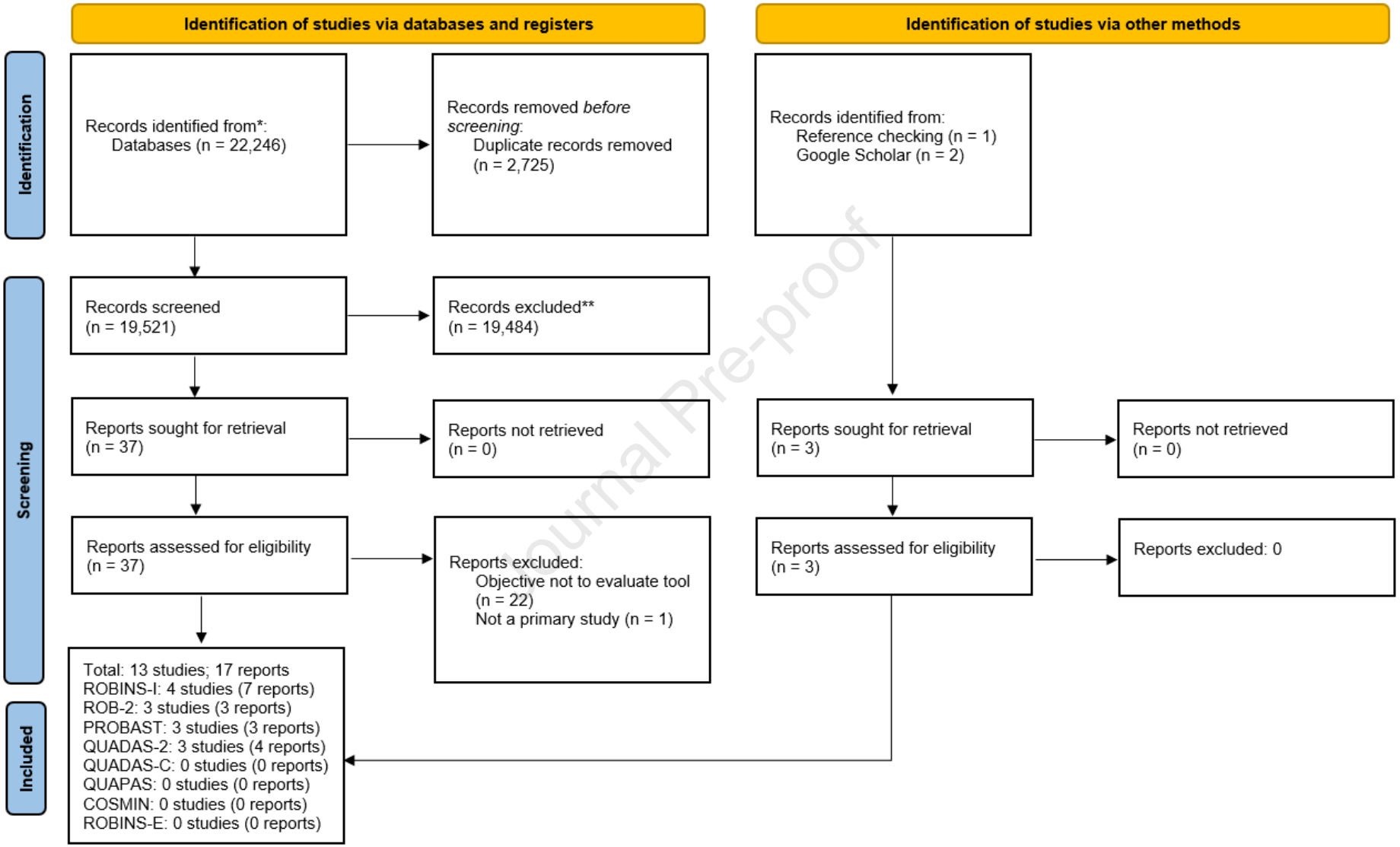

Figure 1 PRISMA diagram showing study selection process for all included risk of bias tools

3.2. Characteristics of included evaluation studies

Characteristics of the 13 included evaluation studies are presented in supplementary material section 3. Studies were conducted by first authors in Australia (1),(24) Brazil (1),(25) Canada (1),(26) China (1),(27) Germany (1),(28) Italy (3; led by the same first author),(29-31) Spain (1),(32) the Netherlands (2),(33, 34) and the United Kingdom (2).(35, 36)

Most studies focused on one clinical topic area, including: cardiology, physical therapy, Alzheimer's disease, uterine cervix examination, cannabis/ cannabinoids for multiple sclerosis, melanoma, independence in older people, and the association between depression and risk of stroke. One study sampled studies focusing on three topics: vaccines, opiate abuse and rehabilitation.(31) Three studies did not restrict to specific topics - one evaluated RoB2,(29) one PROBAST,(34) and one ROBINS-I.(27)

Sampling methodology varied across evaluation studies: six sourced studies from one or more systematic reviews known to, or in production by, the author team conducting the evaluation study;(24, 30-32, 36, 37) three studies searched databases to identify published systematic reviews from which to sample primary studies;(25, 27, 34) three searched databases to identify published studies/ models;(26, 29, 33) and one sampled all melanoma risk prediction studies published until 2021.(28)

In most of the included evaluation studies, two or more individuals independently applied the RoB tool. In one evaluation study, one person completed the RoB assessment and it was checked by another person.(37) In two studies, the authors appraised RoB assessments conducted in other studies: one sampled systematic reviews that had used PROBAST(34) and one sampled meta-analyses that had used QUADAS-2.(25)

Where reported, all study authors used published guidance associated with the tool. One study was conducted as part of the Cochrane RoB2 pilot project, therefore the authors were in contact with the tool developers.(30) Three studies involved the creation and use of extra project-specific guidance about RoB assessment.(26, 28, 30)

The amount of information reported about the assessors' background and experience varied. Most had research experience but had not used the RoB tool under evaluation. Four studies reported having piloted the tool before beginning the study.(24, 27, 32, 37) Two studies reported that assessors had clinical knowledge relevant to the topic area.(24, 31) In other studies it was either not reported or some assessors had clinical backgrounds but it was not clear if it was relevant to the assessment topic. Cross checking the author lists of the included studies with the tool developers of the included RoB tools revealed that none of the evaluation studies were conducted by tool developers.

3.3. Findings from included evaluation studies

Table 2 presents common findings across the RoB tools for which we identified evaluation studies. Study findings should be interpreted in light of the study characteristics previously summarised and tabulated in supplementary material section 3, as the studies varied in clinical topic area, number of individuals using the tool, methodology and user background. Further detail, including user agreement at the domain- and signalling-level and all coded study findings by theme for each tool, is provided in supplementary material section 4 and section 5. Additionally, as previously noted, an overview of the included risk of bias tools, and the similarities and differences between them, can be found in supplementary material section 2.

Table 2 Findings common to more than one RoB tool

Theme	Key finding	Tool	Details from studies
Challenging items	The domain "deviations from intended interventions" is challenging to complete	ROBINS-I	• Challenging to interpret the signalling questions in this domain.(31) • Most time-consuming domain; struggled to decipher if an intervention was a cointervention, whether the cointervention was important and whether it resulted in substantial imbalances between groups.(27)
		RoB2	• Barriers to assessing domain included difficulty in deciding when deviations from the intended intervention did not reflect usual practice and whether they were likely to have affected the outcome, new terminology and approach used in the tool, poor clarity of the guidance, and lack of subject matter and statistics knowledge.(29) • Uncertainty regarding what evidence should be considered to indicate "probably no" to SQ2.3 which asks whether there are deviations from the intended intervention that arose because of the trial context.(36)
	Guidance lacks clarity	PROBAST	• Guidance leaves questions open to interpretation which complicates assessment.(33)
		ROBINS-I	• Lengthy detailed guidance may contribute to limited clarity in overall bias assessment.(26) • Insufficient instructions in guidance on decision trees of conditional signalling questions when the answer to the previous question is "no information" may contribute to poor domain-level agreement between assessors.(31) • Clarity of instructions and items in the tool rated as "poor".(32)
		RoB2	• Lack of clarity in the guidance concerning terminology and approach introduced since RoB1 contributed to difficulties in assessment. Guidance is an improvement to the original version of the tool, but it is complex and its application is demanding.(29) • Extensive guidance but in some places it lacks specificity sufficient to operationalise it, and some parts are unnecessarily discursive and theoretical with insufficient practical advice for interpreting the signalling questions.(36)
Future use and development of tool – suggestions for tool developers	Add intermediate domain-level answer option between "low" and "high" risk (QUADAS-2; PROBAST)	PROBAST	• Add intermediate option to facilitate "less stringent assessment".(33) • Add intermediate option to help distinguish between RoB in studies.(34)
		QUADAS-2	• Add intermediate option to promote consistency.(37)
	Provide clearer guidance	ROBINS-I	• Suggestions to make guidance clearer included: integrate practical examples of biases, and add more examples from different fields of medicine to clarify the questions and highlight situations where actual bias may arise.(27, 31)
		RoB2	• Refine the guidance with a focus on operationalising the tool e.g.: 1) add specific examples directly related to signalling questions and more examples of "judgement calls" rather than extremes; 2) provide greater emphasis on the application of, and location of dividing lines for,

			each signalling question; 3) move some of the theoretical background and empirical evidence to an appendix; 4) provide more guidance regarding a suitable cut-off point for judging whether nearly all data is available in a study.(36)
		QUADAS-2	• Give specific guidance on what constitutes low/ high domain-level bias.(37) • Give clearer instructions on how to assess studies including multiple index tests, reference standards or pathologies.(24)
	Produce simplified versions of tool	PROBAST	• Shorter version of the tool would be useful, particularly when conducting high volume assessments where the classification into high/ low risk is the main objective. Note: the evaluation study authors created a short version in this study.(33)
		ROBINS-I	• Tool is overcomplicated and a simpler version is needed.(27) • Number of signalling questions should be reduced.(31)
Future use and development of tool – suggestions for tool users	Form a team with relevant experience and knowledge	PROBAST	• Include experienced epidemiologists specialised to the area of research and involve them to create valid decision rules for RoB ratings.(28)
		ROBINS-I	• Form a team containing methodology, statistics and clinical expertise.(31)
		RoB2	• Ensure team has clinical, methodological and statistical knowledge.(29, 30)
		QUADAS-2	• Ensure team includes clinicians to assist with applicability assessment.(37)
	Conduct adequate preparatory work before starting RoB assessments including training on how to use the tool and creating project-specific guidance	PROBAST	• Create customised guidance for the use of the tool specific to the topic of study (and present this in the systematic review report) and undertake disease- and study-type-specific training.(28) • Complete specialised training in the tool before use.(34)
		ROBINS-I	• Undertake training on the use of the tool; have a preliminary discussion on how to apply the tool using a common approach; pilot assessment.(31) • Undertake training in the use of the tool and develop an implementation document containing clear decision rules customised to project, to agree on how the assessments will be conducted before beginning.(26) • Create a list of confounders, including time-varying confounding and cointervention, to save assessment time.(27, 32)
		RoB2	• Undertake formal training on how to use tool and have a preliminary discussion on how the tool will be applied, and pilot tool.(29) • Agree on the validity and appropriateness of outcome measures at protocol stage, create an implementation document tailored to review with instructions on how to answer each signalling question and a list of possible cointerventions that could lead to bias (to help with the "deviation from intended interventions" domain), pilot tool.(30) • Develop specific guidance tailored to the review to help to overcome challenges and ensure consistency across assessments.(36)

Floor and ceiling effects	Presence of floor and/or ceiling effects in the tools	PROBAST	• Ceiling effect identified(34) – Users unable to distinguish between studies of varying poor quality e.g. a study scoring poorly in one domain will have the same overall score as a study scoring poorly in multiple domains.(33, 34)
		ROBINS-I	• If all domains are judged as "moderate", the overall judgement will be "moderate", which may lead to inaccurate results/ cover up a serious study issue. (27)
		QUADAS-2	• Floor and ceiling effects not identified in the tool.(24)
Overall RoB judgement	Crude categorisation of overall RoB	PROBAST	• Concerns about overall RoB due to stringent assessment and lack of "intermediate" answer option.(33, 34)
		ROBINS-I	• Discriminative ability of the tool is limited by the "crude" categorisation of overall RoB as "low", "moderate", "serious", or "critical".(27)
	Tailor how signalling questions contribute to overall judgement	QUADAS-2	• Useful to tailor how signalling questions within a domain contribute to the overall domain-level RoB judgment. This provides the opportunity to weight signalling questions to topic specific sources of bias.(37)
Overall usability	Tools are complex to use	ROBINS-I	• Tool is too comprehensive to provide a concise critical appraisal.(27)
		RoB2	• Tool and guidance are comprehensive and though it is an improvement on the last version of the tool, it is complex and demanding to apply.(29) • Tool is complex, particularly the signalling questions.(30) • Conducting assessments with RoB2 is a substantial and challenging undertaking.(36)
Time to complete tool	Large time investment to conduct RoB assessments but this varies depending on factors such as the number of outcomes and reports assessed and the individuals conducting assessments. Time to apply tool was mostly found to reduce with use of project-specific guidance and training	PROBAST	• Estimated that applying tool took under one hour.(33) • Before training sessions (but after having read the tool and guidance document), the first 20 studies had eight hours of consensus discussions (2x4hr meetings). Following training/ customised guidance, the next 22 studies had 9 hours of consensus discussions (6x1.5hr meetings).(28)
		ROBINS-I	• Mean time to apply tool (read article and assess relevant outcomes) was 27.8min (SD 12.6).(31) • Mean time to assess RoB and reach a consensus was 48.45min (95%CI 45.61 to 51.29). It reduced by 12.9min (95%CI -16.4 to -9.4) after guidance/ training.(26) • Time to assess a single study (excluding time to read study) reduced from 7 to 3hr as familiarity increased. Time to reach consensus reduced from 40min to 14min.(27)
		RoB2	• Mean time to apply tool (one outcome per study; not using Excel tool; not including consensus decisions) was 28 minutes (standard deviation (SD) 13.4).(29) • Mean time to apply tool (multiple outcomes per study; not reported what was included) was 168.5 minutes (SD: 68.7). This reduced to 41 minutes (SD: 18.39) when using an implementation document, which itself took 40 hours over the course of three months to develop.(30)

			• Mean time for an individual assessment of a study (peer reviewer, with multiple results and reports) was 127min (2h 7m; SD 67) and 54min (SD 43) for the consensus meeting. In total (for two complete individual assessments and a consensus meeting), the resource for the overall process was 358min (SD 183). For 99 studies, it took 35,472min of worktime (591.2 person-hours or 2.1 months of 2 x FTE work) including each individual assessment and two people in a consensus meeting (5h 58m per study, 47m per result). Number of results and reports per study increased the time to conduct assessments and the former also increased time for consensus meetings. There was substantial variation between individual reviewers and experience reduced time taken to conduct individual assessments and consensus meetings. For two reviewers who had already previously assessed at least 25 studies with RoB2, the overall worktime to assess 32 studies was 5hr 15min per study or 44min per result on average. Regression analysis estimated this to be 178 minutes per study (2h 58m; 95% CI 139 to 218) plus 19 minutes per result (95% CI 15 to 24) (adjusted R2 .73).(36)
		QUADAS-2	• More time consuming than original version of the tool because more free text was required, but this was helpful for forming overall judgements.(37)
User agreement – overall RoB	Low inter-rater agreement for overall RoB in QUADAS-2 and RoB2; varying agreement for PROBAST and ROBINS-I	PROBAST	• 90% agreement between two assessors before final consensus meeting (Kappa 0.33).(33) • Poor agreement before customised training (mean pairwise AC1: 0.098/ mean pairwise Cohen's k: 0.132, multi-rater AC1: 0.071), then moderate after training (mean pairwise AC1: 0.474 / mean pairwise Cohen's k: 0.261, multi-rater AC1: 0.476).(28)
		ROBINS-I	• Moderate agreement (PABAK 0.57).(32) • Slight agreement (Kappa 0.06 (95%CI 0.001-0.12)).(31) • Substantial agreement (AC1 statistic 95%CI 0.61 (0.42-0.79)).(27) • Poor agreement before implementation document (AC1 statistic 95%CI 0.00 (0.00-0.18)), then fair agreement after implementation document used (AC1 statistic 95%CI 0.38 (0.18-0.58)).(26)
		RoB2	• Slight agreement (Kappa 0.16, 95%CI 0.08-0.24).(29) • No agreement (Kappa -0.15) before implementation document used, then moderate agreement (Kappa 0.42) after implementation document used.(30)
		QUADAS-2	• Poor inter-tester reliability of summary scores (ICC = 0.36; 95%CI 0.08-0.59).(24)

RoB: Risk of bias. Note: Study characteristics including how many people used the tool, their experience and topics of studies assessed are outlined in supplementary material section 3.

We identified common findings across tools for 6 of 14 themes (see Table 2): challenging items, overall RoB judgement across domains, tool usability, time to complete tool, user agreement, and future use and development of tools. Of the other eight themes, six only had findings for the QUADAS-2 tool, limiting comparison across tools, and we identified no findings for the theme "re-organisation of questions".

3.3.1. Specific tool items

The domain "deviations from intended interventions" was found to be difficult to complete in two studies of RoB2(27, 31) and two studies of ROBINS-I(29, 36) (see Table 2).

3.3.2. Design of tool

Tool users raised concerns about tool ceiling effects and the way that overall RoB is calculated within PROBAST(33) (34)and ROBINS-I(27), noting that the tool structure does not allow users to distinguish between studies of varying poor quality, as studies will be rated at high RoB overall, whether they are high risk in one or all domains.

3.3.3. Tool usability

Many studies reported that the RoB tools were challenging and time consuming to use, due to their length and complexity. The time it took to implement the tools varied and depended on the number of outcomes being assessed and on the tasks included in measurement (e.g. including consensus discussions or not). It reduced with the use of project-specific guidance and training.(26, 28, 30)

User agreement for overall RoB varied across tools. QUADAS-2(24) and RoB2(29, 30) had low inter-rater agreement. It varied across studies for PROBAST(28, 33) and ROBINS-I.(26, 27, 31, 32) Inter-rater agreement at the domain- and signalling question- level for each tool is outlined in supplementary material section 5, showing variation between studies.

3.3.4. Training and guidance

Five studies of PROBAST, ROBINS-I and RoB2 noted that the guidance provided by tool developers lacks clarity.(26, 29, 31-33, 36) Studies called for clearer guidance,(24, 27, 31, 36, 37) and simplified versions of the tools.(27, 31, 33) Three studies recommended that QUADAS-2 and PROBAST should include an intermediate domain-level answer option between low and high risk.(33, 34, 37)

Studies (all tools) commonly posited two pieces of advice for tool users: 1) form a team with methodological, statistical and clinical expertise,(28-31, 37) and 2) conduct preparatory work before beginning RoB assessments, such as training, piloting the tool and creating project-specific RoB guidance.(26-28, 31, 32, 34)

4. Discussion

4.1. Summary of main results

Thirteen studies have evaluated the design and usability of the PROBAST, RoB2, ROBINS-I, and QUADAS-2 RoB tools. These constitute four of eight key RoB tools specified by the LATITUDES Network. No study evaluations were identified for the other tools: QUADAS-C, QUAPAS, ROBINS-E, and the COSMIN RoB checklist. With the exception of the COSMIN RoB checklist, this may because they were created more recently.

We have shown that there are common findings from evaluation studies across RoB tools concerning tool design and usability (e.g. concerns with tool length and complexity, overall bias calculations, and varying user agreement), and recommendations for tool users (e.g. conduct

adequate preparatory work prior to RoB assessment) and tool developers (e.g. produce clearer guidance and simplify tools).

4.2. Strengths and limitations

This paper is the first to provide a review of evaluations of RoB tools, limiting our ability to interpret the findings within the context of other available evidence. Strengths of this review include that it followed a pre-registered protocol and adhered to pre-specified inclusion criteria. The author team includes tool developers and tool users, therefore we had a robust understanding of how the tools are intended to be used. We used objective tool inclusion criteria and members of the team who have been involved in the development of included tools did not undertake screening, data extraction or synthesis. Nonetheless, we acknowledge that our prior experience may have influenced how we interpreted results.

A potential limitation of this review is that studies were only included if they had reported an objective to evaluate one of the included RoB tools within the title or abstract. We may therefore have missed studies that evaluated a tool but not as a main objective. Additionally, we may have missed evaluations by focusing only on published work. In future, it would be useful to supplement this review by interviewing author teams who have used RoB tools and by contacting tool developers for their own tool evaluations. It may also be insightful to explore evaluations of a wider sample of RoB tools. This review focused on eight tools specified as key tools by the LATITUDES Network, however these are only a handful of those available.(4, 6)

Our findings should be interpreted within the context of the characteristics of the included evaluation studies, which varied in methodology, approach to bias assessment, clinical topic area and tool user background. The challenges faced by tool users in the studies may be topic dependant and therefore the generalisability of our findings may be limited. However, we have included a range of clinical topic areas in this review.

We identified limitations in the evaluation methodology employed by included studies. Not all evaluations were conducted in "real life" settings: for example, one team received support from tool developers as part of a Cochrane RoB2 pilot and some other studies assessed only one outcome per study, which may not often be the case in practice. One study that evaluated ROBINS-I used a sample of cohort studies of exposure, rather than non-randomised studies of interventions, for which the tool is intended to be used.(32) Additionally, experience and topic expertise varied across evaluation teams making generalisability difficult to judge, and the design of the evaluations differed between studies. Some studies focused on one clinical topic area and some spanned multiple areas, with one noting that the generalisability of its findings was limited due to the inclusion of 19 clinical specialities.(29) However, despite this variation, themes emerged across included studies, adding confidence to the validity and applicability of our findings.

4.3. Implications

The findings of this review will be useful to individuals planning to develop, update, use or evaluate a RoB tool. Tool developers may benefit from incorporating the findings of existing evaluations into tool updates, as is currently being done by members of the present author team when updating QUADAS-2 into QUADAS-3. Tool developers may also find it helpful to appraise the similarities and differences between these widely used tools, with a view to aligning (where relevant) with other tools, to support reviewers who are likely to use more than one tool. We have shown that there are common themes across tools that could be addressed as part of aligning key features in design and support/guidance available for tool users. Additionally, as this review has shown that the process of using RoB tools can be complicated and RoB tool guidance can lack clarity, tool users may benefit

from reading our summary of the recommendations for use of the specific tools, given by the authors of the included evaluation studies (e.g. to develop review-specific guidance to assist the team in interpreting the generic RoB tool guidance). Finally, anyone seeking to evaluate a RoB tool in the future may be keen to learn from the methodology used in existing evaluation studies and the limitations of previous work.

To support future evaluations of RoB tools, guidance on good practice for tool evaluation is needed. Learning from the limitations outlined in existing evaluation studies, we suggest a few starting points to consider within tool evaluations. Firstly, consider clinical topic area. Whilst it is useful to apply the tool to studies from a range of topics to improve the generalisability of results, it is also important to reduce noise that is not related to tool features and usability. Therefore, it may be useful for authors of evaluation studies to limit to one or a few topics. Secondly, consider the clinical and RoB assessment experience of the research team and try to mirror the "real life" use of the tool as much as possible. Lastly, consider which outcomes and measures are most helpful to the evaluation. A core set of outcomes could be developed to improve uniformity across evaluation studies, thus improving our ability to compare study findings.

5. Conclusions

Studies have evaluated the design and usability of RoB tools, highlighting common challenges and recommendations for tool use and development. These findings may be useful for people using, developing or evaluating tools. Guidance is needed to encourage appropriate design and implementation of RoB tool evaluation studies. Such guidance could help to enhance the quality of future evaluations of RoB tools.

Acknowledgements

We thank Amanda Owen-Smith for providing guidance concerning qualitative analysis.

Author contributions

Eve Tomlinson: Conceptualization, Methodology, Project administration, Formal analysis, Investigation, Visualization, Writing - original draft, Writing - review & editing.
Chris Cooper: Conceptualization, Methodology, Writing - review & editing.
Clare Davenport: Conceptualization, Methodology, Supervision, Writing - review & editing.
Anne W.S. Rutjes: Conceptualization, Methodology, Supervision, Writing - review & editing.
Mariska Leeflang: Conceptualization, Methodology, Supervision, Writing - review & editing.
Sue Mallett: Conceptualization, Methodology, Supervision, Writing - review & editing.
Penny Whiting: Conceptualization, Methodology, Supervision, Writing - review & editing.

Declaration of interests

Eve Tomlinson: Since working on this review, Eve has joined the QUADAS-3 steering group.
Chris Cooper: None.
Clare Davenport: Clare Davenport is a member of the Diagnostic test Accuracy Editorial Team and is an author on the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Studies. Clare was a member of the QUADAS-2 working group and a member of the QUADAS-C steering group. She is currently a member of the QUADAS-3 steering group.
Anne W.S. Rutjes: Anne is regularly hired by the Federal office of Public Health in Bern, Switzerland as a methodological consultant for Health Technology Assessment reports and receives teaching fees from the academic sector on topics related to research designs, risk of bias tools and reporting

tools. Anne was an editor for the Cochrane Dementia and Cognitive Improvement group up to March 2023 and is an author on the Cochrane Handbook for Systematic Reviews on Diagnostic Test Accuracy Studies. She was a panel member of PROBAST and various COSMIN tools, part of the steering group of QUADAS and QUADAS-2, is part of the core group developing a RoB tool for prevalence studies and one developing a RoB and reporting tool for use in studies evaluating the incidence of adverse events detected with record review methods.
Mariska Leeflang: Mariska Leeflang is an author of the following risk of bias tools: QUADAS-C, QUADAS-2 and PROBAST. Mariska is an editor of the Cochrane Diagnostic Test Accuracy Handbook.
Sue Mallett: Sue Mallett is an author of the following risk of bias tools: QUADAS-C, QUADAS-2 and PROBAST. She is also on the leadership board for the LATITUDES Network. No payments are associated with this.
Penny Whiting: Penny is an author on the Cochrane Handbook for Systematic Reviews on Diagnostic Test Accuracy Studies. She led the development of QUADAS and QUADAS-2, and was a steering group member for the development of PROBAST. She contributed to the development of RoB 2 and ROBINS-I. She is also on the leadership board for the LATITUDES Network.

Role of the funding source

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

Data availability

The full summarised data is provided in supplementary material. A data extraction sheet containing the verbatim quotes extracted from included studies is available upon reasonable request.

References

1. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions: John Wiley & Sons; 2019.
2. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. bmj. 2021;372.
3. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. Journal of clinical epidemiology. 2011;64(4):383-94.
4. Ma L-L, Wang Y-Y, Yang Z-H, Huang D, Weng H, Zeng X-T. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? Military Medical Research. 2020;7(1):1-11.
5. Otzen T, Manterola C, Mora M, Quiroz G, Salazar P, García N. Statements, Recommendations, Proposals, Guidelines, Checklists and Scales Available for Reporting Results in Biomedical Research and Quality of Conduct. A Systematic Review. International Journal of Morphology. 2020;38(3).
6. Network L. LATITUDES Network 2023 [Available from: https://www.latitudes-network.org/.
7. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36.
8. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. bmj. 2019;366.
9. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. bmj. 2016;355.

10. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Annals of internal medicine. 2019;170(1):51-8.
11. Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF, et al. QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. Annals of internal medicine. 2021;174(11):1592-9.
12. Lee J, Mulder F, Leeflang M, Wolff R, Whiting P, Bossuyt PM. QUAPAS: an adaptation of the QUADAS-2 tool to assess prognostic accuracy studies. Annals of Internal Medicine. 2022;175(7):1010-8.
13. ROBINS-E Development Group (Higgins J MR, Rooney A, Taylor K, Thayer K, Silva R, Lemeris C, Akl A, Arroyave W, Bateson T, Berkman N, Demers P, Forastiere F, Glenn B, Hróbjartsson A, Kirrane E, LaKind J, Luben T, Lunn R, McAleenan A, McGuinness L, Meerpohl J, Mehta S, Nachman R, Obbagy J, O'Connor A, Radke E, Savović J, Schubauer-Berigan M, Schwingl P, Schunemann H, Shea B, Steenland K, Stewart T, Straif K, Tilling K, Verbeek V, Vermeulen R, Viswanathan M, Zahm S, Sterne J). . Risk Of Bias In Non- randomized Studies - of Exposure (ROBINS-E). 2023 [Available from: https://www.riskofbias.info/welcome/robins-e-tool. .
14. Mokkink LB, De Vet HC, Prinsen CA, Patrick DL, Alonso J, Bouter LM, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Quality of Life Research. 2018;27:1171-9.
15. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. Circulation. 2015;131(2):211-9.
16. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Journal of Pharmacology and pharmacotherapeutics. 2010;1(2):100-7.
17. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet. 2007;370(9596):1453-7.
18. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Radiology. 2015;277(3):826-32.
19. Moher D, Jones A, Lepage L, Group C, Group C. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. Jama. 2001;285(15):1992-5.
20. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Systematic reviews. 2021;10(1):1-11.
21. Tomlinson E. Risk of bias tools: a systematic review of usability 2023 [Available from: https://osf.io/jf3xp/.
22. Cooper C, Booth A, Britten N, Garside R. A comparison of results of empirical studies of supplementary search techniques and recommendations in review methodology handbooks: a methodological review. Systematic reviews. 2017;6(1):1-16.
23. NVivo. 14 ed.
24. Kaizik MA, Garcia AN, Hancock MJ, Herbert RD. Measurement properties of quality assessment tools for studies of diagnostic accuracy. Braz J Phys Ther. 2020;24(2):177-84.
25. Venazzi A, Swardfager W, Lam B, Siqueira JO, Herrmann N, Cogo-Moreira H. Validity of the QUADAS-2 in Assessing Risk of Bias in Alzheimer's Disease Diagnostic Accuracy Studies. Front Psychiatry. 2018;9:221.
26. Jeyaraman MM, Robson RC, Copstein L, Al-Yousif N, Pollock M, Xia J, et al. Customized guidance/training improved the psychometric properties of methodologically rigorous risk of bias instruments for non-randomized studies. J Clin Epidemiol. 2021;136:157-67.

27. Zhang Y, Huang L, Wang D, Ren P, Hong Q, Kang D. The ROBINS-I and the NOS had similar reliability but differed in applicability: A random sampling observational studies of systematic reviews/meta-analysis. J Evid Based Med. 2021;14(2):112-22.
28. Kaiser I, Pfahlberg AB, Mathes S, Uter W, Diehl K, Steeb T, et al. Inter-Rater Agreement in Assessing Risk of Bias in Melanoma Prediction Studies Using the Prediction Model Risk of Bias Assessment Tool (PROBAST): Results from a Controlled Experiment on the Effect of Specific Rater Training. Journal of Clinical Medicine. 2023;12(5):1976.
29. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol. 2020;126:37-44.
30. Minozzi S, Dwan K, Borrelli F, Filippini G. Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) improved with the use of implementation instruction. J Clin Epidemiol. 2022;141:99-105.
31. Minozzi S, Cinquini M, Gianola S, Castellini G, Gerardi C, Banzi R. Risk of bias in nonrandomized studies of interventions showed low inter-rater reliability and challenges in its application. J Clin Epidemiol. 2019;112:28-35.
32. Losilla JM, Oliveras I, Marin-Garcia JA, Vives J. Three risk of bias tools lead to opposite conclusions in observational research synthesis. J Clin Epidemiol. 2018;101:61-72.
33. Venema E, Wessler BS, Paulus JK, Salah R, Raman G, Leung LY, et al. Large-scale validation of the prediction model risk of bias assessment Tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. J Clin Epidemiol. 2021;138:32-9.
34. Langenhuijsen LF, Janse RJ, Venema E, Kent DM, van Diepen M, Dekker FW, et al. Systematic meta-review of prediction studies demonstrates stable trends in bias and low PROBAST inter-rater agreement. Journal of Clinical Epidemiology. 2023.
35. Wade R, Spackman E, Corbett M, Walker S, Light K, Naik R, et al. Adjunctive colposcopy technologies for examination of the uterine cervix - DySIS, LuViva Advanced Cervical Scan and Niris Imaging System: a systematic review and economic evaluation. Health Technology Assessment. 2013;17(8):1-+.
36. Crocker TF, Lam N, Jordão M, Brundle C, Prescott M, Forster A, et al. Risk-of-bias assessment using RoB2 was useful but challenging and resource-intensive: observations from a systematic review. Journal of Clinical Epidemiology. 2023.
37. Wade R, Corbett M, Eastwood A. Quality assessment of comparative diagnostic accuracy studies: our experience using a modified version of the QUADAS-2 tool. Res Synth Methods. 2013;4(3):280-6.

What is new?

Key findings

- Thirteen studies have evaluated the design and usability of risk of bias (RoB) tools: PROBAST, RoB2, ROBINS-I, QUADAS-2. The studies highlight common challenges with tool implementation and provide recommendations for tool use and development. We identified no primary studies evaluating QUADAS-C, QUAPAS, ROBINS-E, or the COSMIN RoB checklist.
- Some of the existing tool evaluation studies have methodological limitations restricting the generalisability of their findings.

What this adds to what is known

- This is the first review to summarise findings from studies evaluating the design and usability of RoB tools.

What is the implication, what should change now

- Our findings will be helpful to tool users and developers who may benefit from other experiences of applying RoB tools and recommendations for improving the RoB assessment process. Anyone planning to evaluate a RoB tool may also benefit from our review of the current literature on tool evaluation, evaluation methodology used and limitations of existing research.
- Guidance is necessary to ensure appropriate design and implementation of RoB tool evaluation studies in future.

Declaration of Interest statement

Eve Tomlinson: Since working on this review, Eve has joined the QUADAS-3 steering group.
Chris Cooper: None.
Clare Davenport: Clare Davenport is a member of the Diagnostic test Accuracy Editorial Team and is an author on the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Studies. Clare was a member of the QUADAS-2 working group and a member of the QUADAS-C steering group. She is currently a member of the QUADAS-3 steering group.
Anne W.S. Rutjes: Anne is regularly hired by the Federal office of Public Health in Bern, Switzerland as a methodological consultant for Health Technology Assessment reports and receives teaching fees from the academic sector on topics related to research designs, risk of bias tools and reporting tools. Anne was an editor for the Cochrane Dementia and Cognitive Improvement group up to March 2023 and is an author on the Cochrane Handbook for Systematic Reviews on Diagnostic Test Accuracy Studies. She was a panel member of PROBAST and various COSMIN tools, part of the steering group of QUADAS and QUADAS-2, is part of the core group developing a RoB tool for prevalence studies and one developing a RoB and reporting tool for use in studies evaluating the incidence of adverse events detected with record review methods.
Mariska Leeflang: Mariska Leeflang is an author of the following risk of bias tools: QUADAS-C, QUADAS-2 and PROBAST. Mariska is an editor of the Cochrane Diagnostic Test Accuracy Handbook.
Sue Mallett: Sue Mallett is an author of the following risk of bias tools: QUADAS-C, QUADAS-2 and PROBAST. She is also on the leadership board for the LATITUDES Network. No payments are associated with this.
Penny Whiting: Penny is an author on the Cochrane Handbook for Systematic Reviews on Diagnostic Test Accuracy Studies. She led the development of QUADAS and QUADAS-2, and was a steering group member for the development of PROBAST. She contributed to the development of RoB 2 and ROBINS-I. She is also on the leadership board for the LATITUDES Network.