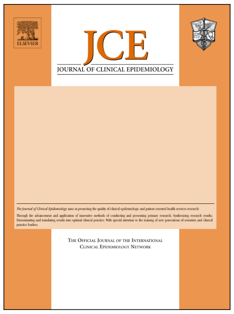

Piloting QUADAS-3: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies

Eve Tomlinson, Bada Yang, Clare F. Davenport, Anne W.S. Rutjes, Sue Mallett, Penny F. Whiting

PII: S0895-4356(25)00316-6

DOI: https://doi.org/10.1016/j.jclinepi.2025.111983

Reference: JCE 111983

To appear in: *Journal of Clinical Epidemiology*

Received Date: 13 June 2025

Revised Date: 2 September 2025

Accepted Date: 14 September 2025

Please cite this article as: Tomlinson E, Yang B, Davenport CF, Rutjes AW, Mallett S, Whiting PF, Piloting QUADAS-3: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies, *Journal of Clinical Epidemiology* (2025), doi: https://doi.org/10.1016/j.jclinepi.2025.111983.

Piloting QUADAS-3: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies

Eve Tomlinson[a], Bada Yang[b], Clare F Davenport[c,d], Anne WS Rutjes[e], Sue Mallett[f], Penny F Whiting[a]

[a] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[b] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

[c] Department of Applied Health Research, University of Birmingham, Birmingham, B15 2TT, UK

[d] NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham B15 2TT, UK

[e] UniCamillus - Saint Camillus International University of Health Science International Medical University, Rome, Italy

[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: QUADAS-2 is the most widely used tool for evaluating risk of bias and applicability concerns in diagnostic test accuracy studies within systematic reviews. QUADAS-2 has recently been updated to a new version, named QUADAS-3. This paper outlines the piloting process undertaken as part of the development of QUADAS-3.

Study design and Setting: Multi-stage piloting: 1) piloting by the QUADAS-3 steering group on a set of five journal papers, 2) piloting workshop at the Global Evidence Summit attended by 16 participants, 3) think aloud interviews with seven researchers who piloted the tool whilst verbalising their thoughts, 4) piloting in five ongoing or completed systematic reviews by seven review authors who provided feedback in an online survey.

Results: Feedback on the tool was generally positive across the four piloting stages. Participants appreciated the structure of the tool, assessment at the estimate level, and the introduction of a framework to define the ideal test accuracy trial. Participants provided suggestions for improvement to the structure and wording of the tool; this led to key changes including the insertion of descriptive prompts within the QUADAS-3 domains, a section at the beginning of the tool to outline the tool's phases and when they should be completed, and clearer wording throughout the tool. Participants also identified areas where further guidance is required for users, including development of worked examples, which will be covered in the associated QUADAS-3 guidance document.

Conclusion: Extensive piloting has ensured that feedback from potential users has been integrated into the development of QUADAS-3.

Keywords: QUADAS-3, Diagnostic Test Accuracy, Risk of Bias, Applicability, Systematic Review, Piloting

Running title: QUADAS-3 piloting

Word count: 3252

Highlights

- This paper outlines the piloting process that informed the development of QUADAS-3 (an updated version of the QUADAS-2 tool) for risk of bias and applicability assessment of diagnostic test accuracy studies in systematic reviews.
- The tool underwent four piloting stages, with revisions made after each stage: 1) steering group piloting on five papers, 2) Global Evidence Summit piloting workshop, 3) think aloud interviews, 4) piloting in systematic reviews.
- Participants' feedback was largely positive, with particular appreciation for the tool's new structure and the shift from study-level to estimate-level assessment.
- Participants provided suggestions to improve wording and structure, which informed key changes to the QUADAS-3 tool.

Plain language summary

What is the problem?

Doctors often use tests to find out if a person has a certain condition. It is important that these tests can correctly tell people who have the condition from those who do not. This is called test accuracy.

Diagnostic reviews are a type of research that bring together results from different studies about the accuracy of a test.

In diagnostic reviews, researchers need to check whether the studies they include are reliable. They also need to check whether the studies match the question the review looks at. Researchers can use a tool called QUADAS-2 to do this. QUADAS-2 was made in 2011 and is now out of date.

We have made a new version called QUADAS-3. In this paper, we explain how we tested the QUADAS-3 tool before making the final version.

What did we do?

We tested the tool in four steps:

1. The people making the tool used it on five research papers
2. It was tested by 16 people at a conference
3. Seven researchers said their thoughts out loud while using the tool and we gathered their feedback
4. Seven researchers used the tool in real reviews. They gave feedback in an online survey

We updated the tool after each step.

What did we find?

Most people liked the QUADAS-3 tool. They also told us ways to make it better. We made changes in response to people’s comments. This led to the final version of QUADAS-3.

1. Introduction

QUADAS-2(1) is the most commonly used and recommended tool for the assessment of risk of bias and applicability concerns in diagnostic test accuracy (DTA) studies included in systematic reviews.(2, 3) The tool consists of four domains: Patient Selection, Index Test, Reference Standard, and Flow and Timing. Risk of bias is assessed for all four domains and applicability is assessed for the first three.

QUADAS-2 was published in 2011.(1) Although feedback on the tool has been largely positive, a number of improvements to the tool have been suggested by users of the tool, both anecdotally and within published tool evaluations.(4) Suggestions include alignment of the tool with more recently developed risk of bias tools for other study designs, improvement in the clarity of the tool guidance, and the incorporation of minor changes made to the version presented in the latest edition of the Cochrane DTA Handbook.(3)

The QUADAS-2 tool has been updated to a new version, named QUADAS-3 (under review). The updated tool incorporates a number of changes including: introduction of the concept of the ideal test accuracy trial; assessment of quality at the estimate level rather than study level; introduction of rationale for judgements, and introduction of guidance to support an *overall* judgement of risk of bias and concerns regarding applicability. We have also made some changes to domains and signalling questions, including amendment of the answer options to signalling questions from "yes", "no," or "unclear" to "yes", "probably yes", "probably no", "no", or "no information" and a change to domain-level answer options from "low", "high", or "unclear" to "low", "high", or "insufficient information". Inclusion of the "probably" option aims to encourage users to answer question even when there may be limited information reported in the paper to allow them to be confident in their answers. A "probably yes" is interpreted in the same way as "yes" when arriving at domain level judgements – if all signalling questions are answered as "yes" or "probably yes" then the domain should be judged as low risk of bias. If one or more signalling questions are answered as "no" or "probably no" then reviewers should use their judgement to determine whether the issue flagged by the signalling question may have introduced bias into the study.

A criticism of QUADAS-2 was of the limited piloting of the tool before publication. Therefore, an important step in the development of QUADAS-3 was a multi-staged piloting process. In this paper, we outline this process and summarise key changes arising from each stage of piloting.

2. Methods

We undertook four stages of piloting. Table 1 provides an overview of these stages and the main changes made to the tool after each stage. We discuss this narratively below.

Further information about each stage, including the versions of the tool we used, are provided in supplementary material.

Table 1 Overview of piloting stages

Stage (tool version)	**Details**	**Main changes made after piloting stage***
1. Steering group piloting (v.0.6)	Two rounds of piloting within the steering group: 1) tool v0.6 piloted on 5 study reports with each one assessed by 3 to 5 group members who provided feedback, 2) tool revised based on feedback and v0.7 sent to all steering group members for any further comments on the tool as a whole	• Simplified Phase 1 "definition of synthesis question" table • Amendments to wording in Phase 2 tables • Added field to specify study ID • In "characteristics of numerical accuracy estimates" table in Phase 4, "study group" changed to "population" and added row for "analysis" • Added note in Phase 5 to clarify that if patients dropped out/excluded from study because they did not receive the index test and/or reference standard then this should be handled in the analysis domain rather than the participants domain • First signalling question in participants domain reworded from "single group" to "single gate" • Removed "green flags" section at end of tool which aimed to identify additional generic features of study design and conduct that indicate good research practice has been followed
2. Global Evidence Summit workshop (v0.8)	Piloting workshop attended by 16 participants with varying experience with QUADAS-2	• In "framework to define ideal test accuracy trial" table in Phase 2, reworded text to improve clarity e.g. defined "prospective design" • In Phase 2 "definition of the ideal trial" table, merged target condition and reference standard rows to form "definition of the target condition" row • Reduced text before Phase 4 table and Phase 5 assessment • Added prompts to domain descriptive boxes
3. Think-aloud interviews (v0.11)	Seven researchers with varying experience with QUADAS-2 piloted QUADAS-3 whilst verbalising their thoughts in a think-aloud interview	• Added a section before Phase 1 to outline the phases of the tool and to instruct the user to read the guidance ("Explanation and Elaboration") document • In Phase 1 explained "sufficient detail" in "definition of synthesis question" • In Phase 2 amended wording in "framework to define ideal test accuracy trial" table to improve clarity • In Phase 4 table clarified what is meant by "domains to be assessed" and revised accompanying text • In Phase 5 improved clarity in the instructions concerning answer options and emphasised the need to complete an assessment for each estimate • Moved the following Phase 5 note into the description box of the Participants domain: "If patients dropped out/excluded from study because they did not receive the index test and/or reference standard then this should be handled in the analysis domain" • Added the signalling question and domain-level answer options to the domain answer boxes

		• Edited the domain-level risk of bias judgement box to say “risk that the selection of participants has introduced bias” (previously “could the selection of participants have introduced bias?”) • Changed the domain-level applicability judgement box to say “concern that the included participants do not match those in the ideal trial” (previously “is there concern….?”) • Amended applicability section heading from “concerns regarding applicability” to “concerns regarding applicability to the systematic review synthesis question” • Changed title of domain “assessment of the target condition” to “target condition” • Amendments to some signalling questions, descriptive boxes and the overall risk of bias section
4. Piloting in systematic reviews (v0.13)	Seven review authors piloted QUADAS-3 in ongoing or completed reviews and provided feedback via structured questionnaire	• Added Phase 6 to the first table in the tool that outlines the phases and when to complete them • Reinstated Study ID field (had been removed in an earlier version) • Added “if applicable” after synthesis question 2 in Phase 1 • Removed a box from applicability assessment so there is only one box for description and rationale

*See supplementary material for details about each piloting stage, including the versions of the tool used at each stage. Phase 1: state the systematic review synthesis questions; Phase 2: define the ideal test accuracy trial for each synthesis; Phase 3: flow diagram; Phase 4: identify the numerical accuracy estimates to assess for risk of bias and applicability; Phase 5: risk of bias and applicability assessment; Phase 6: overall judgement.

2.1 Steering group piloting

The QUADAS-3 steering group consisted of 13 members. Seven of these 13 members formed the core group, who met regularly to advance tool development.

In this first stage, conducted in July 2024, members of the steering group piloted the tool (v0.6) (Appendix 1) on a set of five study reports. The reports were selected to cover a broad range of target conditions, index test, and reference standard types. The articles included some challenging issues in terms of risk of bias and applicability assessment (Table 2), such as having multiple recruitment sites, several reference standards, and missing data.

Each study report was assessed independently by between three and five individuals. We defined hypothetical systematic review questions (‘synthesis questions’) and ideal test accuracy trials for each clinical topic area to allow for assessments of applicability. Each of the steering group members shared details of their experience of using the tool and made suggestions for improvements. We used this feedback to develop an updated draft of the tool (v0.7). We shared the updated draft with the steering group for further comments on the tool as a whole. The additional steering group feedback was then incorporated into a further update to the tool (v0.8).

Table 2 Overview of study reports assessed as part of the steering group piloting

Study report	**Population**	**Index test**	**Target condition**	**Reference standard**
Kidd et al. (2022)(5)	Asymptomatic and symptomatic individuals across health care and community settings	Four RT-LAMP assays performed on nasopharyngeal/ oropharyngeal swab and saliva samples	SARS-CoV-2	RT-qPCR
Gasem et al. (2002)(6)	Patients with clinical suspicion of typhoid presenting to hospital	Dipstick assay performed on blood samples	Typhoid and paratyphoid infection	Bone marrow and blood culture
Hollis et al. (2018)(7)	Adults and children with suspected ADHD	QbTest (computer-based continuous performance task)	ADHD	Clinical assessment
McCarthy et al. (2007)(8)	Patients with suspected coronary artery disease	TrueFISP breathold coronary MRI	Coronary artery disease	X-ray angiography
Baraliakos et al. (2020)(9)	People with chronic back pain	Algorithm based on patient questionnaire of symptoms in combination with HLA-B27 blood test	Axial spondyloarthritis	Two rheumatologists, with laboratory analyses including C-reactive protein and imaging (MRI and x-rays)

ADHD: attention deficit hyperactivity disorder. RT-LAMP: reverse transcription loop-mediated isothermal amplification. MRI: magnetic resonance imaging. RT-qPCR: quantitative reverse transcription polymerase chain reaction.

2.2 Global Evidence Summit workshop

The draft of the QUADAS-3 tool created in stage one (v0.8) (Appendix 2) was then piloted in a workshop at the Global Evidence Summit in September 2024. We introduced the draft tool during a short presentation and answered questions from participants. We then provided the tool to all participants together with a study report (Gasem et al.(6)) used in the previous piloting phase (Table 2). The tool had Phase 1 (state the systematic review synthesis question), Phase 2 (define the ideal test accuracy trial for each synthesis question), Phase 3 (flow diagram), and Phase 4 (identify the numerical accuracy estimates to assess) completed. Table 2We assigned participants to four smaller groups and allocated each group to try out one of the QUADAS-3 domains in Phase 5 (risk of bias and applicability assessment). We asked the groups to complete the risk of bias and concerns regarding applicability assessment for their domain. We gathered feedback in a group discussion which was used by the steering group as the basis for subsequent drafts of the tool.

2.3 Think-aloud piloting

Following the Global Evidence Summit workshop, two interim versions of the tool (0.9; 0.10) were developed containing minor adjustments from regular meetings of the core group. A subsequent version of the tool (v0.11), which was further informed by meetings of the core group, then underwent think-aloud piloting. This version and the detailed methods for this piloting stage are outlined in Appendix 3.

We recruited participants known to the core group with different levels of experience of QUADAS-2 to take part in an online think-aloud interview. Participants were provided with the same study used for the Global Evidence Summit workshop,(6) and a template of QUADAS-3 with Phase 1 to 4 completed. Participants were asked to verbalise their thoughts on Phases 1 to 4 and then complete Phase 5 (risk of bias and applicability assessment) and Phase 6 (overall judgement) of the tool, talking through their thoughts as they completed the task. As part of this stage of piloting, we also explored preferences for signalling *questions* versus signalling *statements*, by giving some participants the *questions* version and some the *statements* version. Those that had been given the *questions* version were shown the *statements* version at the end of the piloting and vice versa, to get feedback on which they preferred. Near the end of the interview, participants were also asked to share their views on how the tool compares to QUADAS-2 (for those experienced with this tool) and any further comments. We took notes and a transcript from the interviews. We summarised findings concerning positives of the tool, errors, difficulties and problems, and suggestions for changes.

2.4 Piloting in systematic reviews

A subsequent revision of the tool (v0.13) (Appendix 4) was then piloted in systematic reviews by authors known to the core group. It was either piloted in completed reviews in which QUADAS-2 had been used, or ongoing reviews where the authors had agreed to use QUADAS-3. We invited review authors to share their feedback on their experience of each phase of the tool through a short web-based survey (Appendix 4). Following this stage, we revised the tool and shared it with the steering group for comment before finalisation.

3. Results

Key changes made to the tool following each stage of piloting are summarised in Table 1.

3.1 Steering group piloting

Feedback from steering group members during piloting included the following main areas for improvement: simplification of the "definition of synthesis question" table in Phase 1, addition of a field to state study ID in Phase 3 (flow diagram), and the removal of a section at end of tool which asked the user to identify any "green flags" (additional generic features of study design and conduct that highlight where good research practice has been followed). A key output to come from this piloting stage was a set of model answers to help users see how the tool should be applied in practice. They will be made available on the QUADAS website (https://www.bristol.ac.uk/population-health-sciences/projects/quadas/).

3.2 Global Evidence Summit workshop

Sixteen participants attended the Global Evidence Summit workshop. Thirteen participants had used QUADAS-2; three had not but they were familiar with DTA reviews. A summary of the workshop findings is presented in Appendix 2. Generally, participants liked the tool, appreciated the introduction of the "probably" answer options, and liked that assessment was conducted at the estimate level.

One participant asked for increased clarity in the tool regarding whether a signalling question could be answered "no" and the domain judged "low risk of bias". This was actioned in a later version of the tool, to clearly explain that this is possible. Some participants commented that that the domain-level applicability question (e.g. "Is there concern that the included participants do not match the review question?") is currently answered with "low", "high", or "unclear" and these answers do not suit a question. This was revised in the next piloting stage to change the question to a statement (e.g. "Concern that the included participants do not match those in the ideal test accuracy trial"). Participants at the workshop had no strong feelings as to whether the tool should contain signalling *questions* or *statements*.

Other suggestions for changes to the tool at this stage mainly related to providing further explanation of the signalling questions. For example, some participants asked for clarification regarding what was meant by the "recommended instructions" in the following signalling question "Was the index test conducted and interpreted according to the recommended instructions?". This, and other explanation for signalling questions, will be provided in the associated QUADAS-3 guidance, named the "Explanation and Elaboration" document.

3.3 Think-aloud piloting

Seven participants each took part in a think-aloud interview. All participants were researchers who had varying experience with QUADAS-2 and of the clinical study area (typhoid). Three were native English-language speakers and four were not. Participant characteristics and a summary of the key findings from the think-aloud piloting is provided in Appendix 3. Changes made to the tool at this stage are summarised in Table 1.

As in the previous piloting stage, participants liked the new answer options including “probably”. They also appreciated the consistency with other frequently used risk of bias tools, including ROBINS-I which asks users to define the ideal (“target”) trial.

Several improvements were also suggested by participants. In Phase 1, participants sought clarification around what was meant by “sufficient detail” when asked to specify the synthesis question. We therefore added the following prompt to help users: “Consider including the following components: population, index test, and target condition.”

In Phase 2, participants suggested improvements to the clarity of the wording in the pre-filled table that defines the ideal accuracy trial and the table in which the user defines the ideal trial for each synthesis question (Table 2 and Table 3 in the final tool). In response, we made changes to the wording.

Some participants suggested that there was duplication in the information the user was required to put into their data extraction forms, the Phase 2 table for defining the ideal trial, and the Phase 5 assessment. No changes were made to the tool concerning this point. It is important that this information is stated in each of these phases of the review and QUADAS-3 to ensure a transparent and thorough risk of bias and applicability assessment.

Some of the suggestions that arose in the previous piloting stage were also raised in the think-aloud piloting. For example, in Phase 5, participants felt there should be further clarification around whether you can answer a question with “no” and still judge risk of bias as “low”. Participants asked what the meaning of “recommended instructions” for the index test was. We responded to the first point by making clear in the tool that a signalling question can be answered “no” and the domain still be judged at low risk of bias. We also provided guidance on the meaning of “recommended instructions” within the associated Explanation and Elaboration document.

As in the previous piloting stage, there was no clear preference for either signalling *questions* versus *statements*.

3.4 Piloting in systematic reviews

Seven review authors piloted the tool (v0.13) (Appendix 4) in five DTA systematic reviews. The research questions of these reviews are outlined in Appendix 4. The target conditions assessed in the reviews included: motor seizures, coeliac disease, tuberculosis (two reviews), and non-ST-elevation myocardial infarction. A summary of the feedback from the survey of review authors is provided in Appendix 4.

Comments from the review authors were generally positive, with participants reporting that the tool was clear and helpful. Some participants noted repetition between the rationale and support for judgement in Phase 5 of the tool. This was amended after this piloting stage so that there is only one box for applicability rationale in the final tool. Useful suggestions were also highlighted for the associated QUADAS-3 Explanation and Elaboration document, which is in preparation.

4. Discussion

We have carried out a comprehensive series of piloting exercises as part of the development of the new QUADAS-3 tool. Overall, feedback on the new tool has been positive, with participants in each stage suggesting helpful areas for improvement. Key improvements made to the tool as a result of piloting included: adding an introductory section to help users navigate the six phases of QUADAS-3, consistency in terminology across the tool, clarity of wording and instructions throughout the tool, rephrasing of domain-level bias and applicability items (Phase 5) from questions to statements to provide a clearer link with the answer options, and rephrasing the items about applicability to link directly to the synthesis question (Phase 1) and ideal test accuracy trial (Phase 2) sections of the tool. The feedback from the piloting informed successive versions of the tool and has also been helpful in identifying areas where guidance is required. We have collated these suggestions, which will be used to inform the Explanation and Elaboration document currently in development. This document will provide guidance on each phase of the tool, including explanations of how to approach each signalling question, and how to reach domain-level and overall risk of bias and applicability judgements.

QUADAS-2 had a limited piloting process, therefore the multi-stage piloting process outlined in this paper for QUADAS-3 is a strength of the tool development. A key strength of our approach is the use of different forms of piloting to gather feedback in different ways from a range of stakeholders. The first stage of piloting involved members of the steering group, applying an early draft of the tool to a set of study reports that we had identified as being challenging in terms of risk of bias and concerns regarding applicability. This allowed us to gather feedback at an early stage in the development of the tool from a group of people with extensive experience and understanding of the original QUADAS-2 tool. Subsequent stages included a more diverse group of participants with varying experience with risk-of-bias assessment in general, the QUADAS-2 tool specifically, and of DTA reviews. Two participants from stage three (think-aloud interviews) also contributed to stage four (piloting in DTA reviews) and so were able to see how the tool had changed based on their initial feedback, and to share further feedback after applying the tool to a topic area that they were more familiar with. One paper was used as part of the first 3 stages of piloting,(6) allowing us to gain feedback on the same paper through quite different processes and with a wide range of participants. A potential limitation of this approach is that it might have limited generalizability, as particularly in stage two and three the same one paper was used, which meant we could not gain feedback on the application of the tool in these stages to a broader range of topics. However, the one paper used contained a number of methodological challenges, making it a useful paper to test the tool on. Additionally, in stage one the tool was applied to four different reports, and stage five involved testing the tool in five systematic reviews containing different papers, so multiple topics were covered overall throughout the piloting process.

The use of an iterative, sequential approach meant that we were able to incorporate feedback from each stage, and then use a revised version of the tool in the next stage of piloting. Successive versions of the tool were also informed by extensive discussions among the core group at bi-monthly meetings. At these meetings, issues raised by the piloting were discussed in detail by the group who considered how best to address these changes in both the tool and the Explanation and Elaboration document. Overall, we believe this process is likely to have increased the face validity of the tool.

We did not evaluate inter-rater reliability as part of our piloting process. Whilst inter-rater reliability is often considered an important feature of quality assessment tools and is often included as part of their evaluations, we do not feel this is helpful for tools such as QUADAS-3. These complex tools require some element of subjective judgement that is often limited by poor reporting of primary studies. Review authors therefore need to use their own experience and judgement to determine whether a potential source of bias may have impacted a particular study based on limited information. Different people may make different judgements based on incomplete information often reported in a study. We consider it essential that at least two review authors are involved in the process of applying QUADAS-3, whether independently or with one review author performing the assessment and a second checking this in detail. An important update to QUADAS-3 compared to the QUADAS-2 tool is that we now ask review authors to provide a justification for their domain-level judgements. The process of discussing the issues identified by the assessment helps review authors to achieve consensus on what the key issues are. We consider this a much more helpful and important process than simply considering whether review authors arrive at the same judgement when completing the process independently. If inter-rater reliability is assessed, it would be more helpful to do this between pairs of review authors. This would consider whether the pairs arrive at the same overall decision following independent application and discussion within pairs.

The piloting approach outlined here was designed pragmatically to be completed with limited resources. We did not pre-register the study or pre-specify particular “outcomes” to assess throughout piloting, instead we mostly sought general feedback to the phases and questions within the tool. This worked well for the current project. However, an alternative approach could have been to ask specifically for feedback on certain areas such as clarity of wording, ease of use, and overall coherence, and to have conducted a comparison of the tool before and after piloting-based revisions to more clearly measure "improvement". Future piloting projects could seek to do this, as well as aim to involve a larger and more diverse sample of participants in each piloting stage. No reporting guideline exists for piloting projects such as this, therefore we reported this study according to the journal guidelines. . Results data from each stage of piloting are reported in the supplementary material.

In conclusion, extensive piloting through a series of iterative steps has contributed to the development of QUADAS-3. This process has ensured that the views of potential users of the tool have been incorporated as part of the development process, which we hope has resulted in a more valid and usable tool.

Declaration of generative AI and AI-assisted technologies in the writing process

No generative AI was used to write this manuscript.

CRediT authorship contribution statement

The grid below shows the contribution of each author to the publication:

Author	**ET**	**BY**	**AWSR**	**CFD**	**SM**	**PFW**
Conceptualization	🟩	🟩	🟩	🟩	🟩	🟩
Methodology	🟩				🟩	🟩
Formal analysis	🟩					🟩
Investigation	🟩	🟩	🟩	🟩	🟩	🟩
Resources	🟩					🟩
Data Curation	🟩					🟩
Writing - Original Draft	🟩					🟩
Writing - Review & Editing		🟩	🟩	🟩	🟩	
Visualization	🟩					🟩
Supervision						🟩
Project administration	🟩					
Funding acquisition	🟩					🟩

Declaration of competing interest

All of the authors are part of the QUADAS-3 development group.

Acknowledgments

We would like to thank all those that contributed to each of the piloting stages:

Steering group piloting: Penny Whiting, Eve Tomlinson, Anne Rutjes, Clare Davenport, Bada Yang, Marie Westwood, Yemisi Takwoingi, Johannes Reitsma, Christopher Hyde, Patrick Bossuyt, Jonathan Deeks, Mariska Leeflang, Sue Mallett

Global evidence summit workshop: The 16 participants at the workshop (we did not collect details of participants).

Think aloud piloting: Anneke Damen, Martha Elwenspoek, Ruth Brassington, Rachel James, Catalina Lopez Manzano, Natasha Thorley, Davide Maffei

Piloting in real reviews: Martha Elwenspoek, Ruth Brassington, Marie Westwood, Johanna Ahsberg, Demy Idema, Stephanie Bjerrum, Alia Razid.

Data Availability

Results of each stage of the piloting are reported in the supplementary materials.

Funding

No specific funding was received for this project.

References

1. 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. Ann Intern Med. 2011;155(8):529-36.
2. Ferrante di Ruffano L, Harris IM, Zhelev Z, Davenport C, Mallett S, Peters J, et al. Health technology assessment of diagnostic tests: a state of the art review of methods guidance from international organizations. Int J Technol Assess Health Care. 2023;39(1):e14.
3. Reitsma J, Rutjes A, Whiting P, Yang B, Leeflang M, Bossuyt P, et al. Chapter 8: Assessing risk of bias and applicability. In: Deeks J, Bossuyt P, Leeflang M, Takwoingi Y, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy: Cochrane; 2023.
4. Tomlinson E, Cooper C, Davenport C, Rutjes AWS, Leeflang M, Mallett S, et al. Common challenges and suggestions for risk of bias tool development: a systematic review of methodological studies. J Clin Epidemiol. 2024;171:111370.
5. Kidd SP, Burns D, Armson B, Beggs AD, Howson ELA, Williams A, et al. Reverse-Transcription Loop-Mediated Isothermal Amplification Has High Accuracy for Detecting Severe Acute Respiratory Syndrome Coronavirus 2 in Saliva and Nasopharyngeal/Oropharyngeal Swabs from Asymptomatic and Symptomatic Individuals. J Mol Diagn. 2022;24(4):320-36.
6. Gasem MH, Smits HL, Goris MGA, Dolmans WMV**.** Evaluation of a simple and rapid dipstick assay for the diagnosis of typhoid fever in Indonesia. J Med Microbiol. 2002;51(2):173-7.
7. Hollis C, Hall CL, Guo B, James M, Boadu J, Groom MJ, et al. The impact of a computerised test of attention and activity (QbTest) on diagnostic decision-making in children and young people with suspected attention deficit hyperactivity disorder: single-blind randomised controlled trial. J Child Psychol Psychiatry. 2018;59(12):1298-308.
8. McCarthy RM, Deshpande VS, Beohar N, Meyers SN, Shea SM, Green JD, et al. Three-dimensional breathhold magnetization-prepared TrueFISP: a pilot study for magnetic resonance imaging of the coronary artery disease. Invest Radiol. 2007;42(10):665-70.
9. Baraliakos X, Tsiami S, Redeker I, Tsimopoulos K, Marashi A, Ruetten S, et al. Early recognition of patients with axial spondyloarthritis-evaluation of referral strategies in primary care. Rheumatology (Oxford). 2020;59(12):3845-52.

Highlights

- This paper outlines the piloting process that informed the development of QUADAS-3 (an updated version of the QUADAS-2 tool) for risk of bias and applicability assessment of diagnostic test accuracy studies in systematic reviews.
- The tool underwent four piloting stages, with revisions made after each stage: 1) steering group piloting on five papers, 2) Global Evidence Summit piloting workshop, 3) think aloud interviews, 4) piloting in systematic reviews.
- Participants' feedback was largely positive, with particular appreciation for the tool's new structure and the shift from study-level to estimate-level assessment.
- Participants provided suggestions to improve wording and structure, which informed key changes to the QUADAS-3 tool.

Declaration of competing interest

All of the authors are part of the QUADAS-3 development group.