QUADAS-C

Yang, Bada ; Mallett, Sue; Takwoingi, Yemisi; Davenport, Clare; Hyde, Christopher J; Whiting, Penny F; Deeks, Jon; Leeflang, Mariska Mg; QUADAS-C Group

DOI:
10.7326/M21-2234

Document Version
Peer reviewed version

Citation for published version (Harvard):
Yang, B, Mallett, S, Takwoingi, Y, Davenport, C, Hyde, CJ, Whiting, PF, Deeks, J, Leeflang, MM & QUADAS-C Group 2021, 'QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies', *Annals of internal medicine*, vol. 174, no. 11, pp. 1592-1599. https://doi.org/10.7326/M21-2234

QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies

Bada Yang[a], MD; Sue Mallett[b]*, DPhil; Yemisi Takwoingi[c,d]*, DVM, PhD; Clare F. Davenport[c,d]*, PhD; Christopher J. Hyde[e]*, MD; Penny F. Whiting[f]*, PhD; Jonathan J. Deeks[c,d]*, PhD; and Mariska M.G. Leeflang[a], DVM, PhD for the QUADAS-C Group†

[a] Department of Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, 1105AZ, Amsterdam, The Netherlands

[b] UCL Centre for Medical Imaging, University College London, London, W1W 7TY, UK

[c] Test Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

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

[e] Exeter Test Group, Institute of Health Research, College of Medicine and Health, University of Exeter, Exeter, UK

[f] Population Health Sciences, Bristol Medical School, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK

*These authors contributed equally to this work and the order was determined randomly.

Corresponding author:

Bada Yang

Department of Epidemiology and Data Science, Room J1b-210

Amsterdam UMC, Location AMC

Meibergdreef 9, 1105AZ Amsterdam, The Netherlands

Tel: +31(0)20 5666948

Email: b.d.yang@outlook.com

† The QUADAS-C Group:

Patrick M.M. Bossuyt, PhD (University of Amsterdam), p.m.bossuyt@amsterdamumc.nl
Miriam G. Brazzelli, PhD (University of Aberdeen), m.brazzelli@abdn.ac.uk
Clare F. Davenport*, PhD (University of Birmingham), c.f.davenport@bham.ac.uk
Jonathan J. Deeks*, PhD (University of Birmingham), j.deeks@bham.ac.uk
Jacqueline Dinnes, PhD (University of Birmingham), j.dinnes@bham.ac.uk
Kurinchi S. Gurusamy, MBBS, PhD (University College London), k.gurusamy@ucl.ac.uk
Hayley E. Jones, PhD (University of Bristol), hayley.jones@bristol.ac.uk
Christopher J. Hyde*, MD (University of Exeter), c.j.hyde@exeter.ac.uk
Stefan Lange, MD (Institute for Quality and Efficiency in Health Care, Germany), stefan.lange@iqwig.de
Miranda W. Langendam, PhD (University of Amsterdam), m.w.langendam@amsterdamumc.nl
Mariska M.G. Leeflang*, DVM, PhD (University of Amsterdam), m.m.leeflang@amsterdamumc.nl
Petra Macaskill, PhD (University of Sydney), petra.macaskill@sydney.edu.au
Sue Mallett*, DPhil (University College London), sue.mallett@ucl.ac.uk
Matthew D.F. McInnes, MD, PhD (University of Ottawa), mmcinnes@toh.ca
Johannes B. Reitsma, MD, PhD (University of Utrecht), j.b.reitsma-2@umcutrecht.nl
Anne W.S. Rutjes, PhD (University of Bern), anne.rutjes@ispm.unibe.ch
Alison Sinclair, MD, PhD (Canadian Agency For Drugs And Technologies In Health), alison.sinclair@icloud.com
Yemisi Takwoingi*, DVM, PhD (University of Birmingham), y.takwoingi@bham.ac.uk
Henrica C.W. de Vet, PhD (VU University Amsterdam), hcw.devet@amsterdamumc.nl
Gianni Virgilli, MD (Queen's University Belfast), gianni.virgili@unifi.it
Ros Wade, MSc (University of York), ros.wade@york.ac.uk
Marie E. Westwood, PhD (Kleijnen Systematic Reviews), marie@systematic-reviews.com
Penny F. Whiting*, PhD (University of Bristol), penny.whiting@bristol.ac.uk
Bada Yang*, MD (University of Amsterdam), b.d.yang@outlook.com
*Steering group members

Abstract

Comparative diagnostic test accuracy studies assess and compare the accuracy of two or more tests in the same study. While these studies have the potential to yield reliable evidence regarding comparative accuracy, shortcomings in the design, conduct, and analysis may bias their results. The currently recommended quality assessment tool for diagnostic test accuracy studies, QUADAS-2, is not designed for the assessment of test comparisons.

We developed QUADAS-C as an extension to QUADAS-2 to assess the risk of bias in comparative diagnostic test accuracy studies. Through a four-round Delphi study involving 24 international experts in test evaluation and a face-to-face consensus meeting, we developed an initial version of the tool which was revised and finalized following a pilot study among potential users.

QUADAS-C retains the same four-domain structure of QUADAS-2 (Patient Selection, Index Test, Reference Standard, and Flow and Timing) and is comprised of additional questions to each QUADAS-2 domain. A risk of bias judgment for comparative accuracy requires a risk of bias judgment for the accuracy of each test (resulting from QUADAS-2) and additional criteria specific to test comparisons. Examples of such additional criteria include whether participants either received all index tests or were randomized to index tests, and whether index tests were interpreted blind to the results of other index tests.

QUADAS-C will be useful for systematic reviews of diagnostic test accuracy addressing comparative questions. Furthermore, researchers may use this tool to identify and avoid risk of bias when designing a comparative diagnostic test accuracy study.

Abstract word count: 247
Manuscript word count: 3631

Figures: 1
Tables: 3
References: 26
Running title: QUADAS-C
Keywords: Diagnostic accuracy; Bias; Test comparison; Methodology; Systematic review

Supplement 1: Information contributing to the development of QUADAS-C
Supplement 2: The QUADAS-C tool
Supplement 3: Guidance on how to use QUADAS-C

1. Introduction

Studies of diagnostic test accuracy (DTA) are pivotal in the evaluation of new and existing diagnostic tests and strategies (1). DTA studies can evaluate the accuracy of a single index test, but can also evaluate multiple index tests and compare their accuracy.

Comparison of test accuracy is preferably done in studies directly comparing index tests in the same study, also known as comparative DTA studies (2,3). Comparative DTA studies have the potential to provide rigorous evaluations of test comparisons, unlike comparisons based on separate studies evaluating the accuracy of single tests (4,5). However, like any study, comparative DTA studies need to be evaluated for their validity and applicability before their results can be used for guiding healthcare decisions.

Comparative DTA studies are susceptible to sources of bias that are not captured by the recommended QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) tool for assessing the methodological quality of DTA studies (6). In test comparisons, bias may arise, for instance, when participants receiving index test A represent a different disease spectrum than those receiving index test B, or when results of A are interpreted with knowledge of the results of B and vice versa (7). To account for these and other potential sources of bias, risk of bias assessments need to include additional items specific to comparisons of test accuracy.

An overview of 238 comparative DTA systematic reviews that were published in 2017 showed that risk of bias assessments for test comparisons had been planned or conducted in only two reviews (3). Furthermore, the overview did not identify any risk of bias tools designed for comparative DTA studies.

We developed the QUADAS-C tool (C stands for comparative) for assessing risk of bias in comparative DTA studies. QUADAS-C is not designed as a standalone tool but as an extension to QUADAS-2. QUADAS-C is designed for use in systematic reviews, but investigators can also consult the tool during the planning and design phases of a comparative accuracy study to reduce risk of bias. In this article we explain the development process of QUADAS-C, its scope, and how it should be used.

2. Comparative accuracy questions

We first briefly explain what comparative accuracy questions are and how they differ from questions regarding single test accuracy (Table 1 outlines key differences). Comparative accuracy questions ask how the accuracy of an index test compares to that of another index test for detecting the same target condition. For example, whether Xpert® MTB/RIF Ultra is more sensitive for diagnosing tuberculous

meningitis compared to Xpert® MTB/RIF (8). For a valid comparison, participants receiving index test A should be exchangeable with participants receiving index test B. This can be accomplished by each participant undergoing all index tests (often referred to as a fully paired or within-subject design), or approximated by randomly allocating participants to index tests (randomized design) (2,9,10). Comparative accuracy results can be expressed as absolute or relative differences in sensitivity and specificity, predictive values, area under the curve, or other measures of accuracy including decision analytic measures such as net benefit (11,12). Knowledge about comparative accuracy is important for the selection and recommendation of a test from a number of alternative, competing tests, especially when studies evaluating the effectiveness of test-treatment strategies on patient-important outcomes are absent (13). A key characteristic of comparative accuracy questions is that none of the tests being compared is the reference standard. Rather, the reference standard is a means to verify whether participants have the target condition or not.

Table 1. Differences between single test accuracy and comparative accuracy questions.

	Accuracy of a single test	Comparative accuracy
Health-related question	How accurately can an index test classify individuals who have or do not have the target condition?	How does the accuracy of index test A compare with that of index test B?
Ideal study design	A study in which participants are consecutively or randomly sampled and all undergo a single index test and the reference standard	A study in which participants are consecutively or randomly sampled and: each participant undergoes all index tests and the reference standard (fully paired or within-subject design) or participants are randomly allocated to an index test and all participants receive the reference standard (randomized design)
Summary measures	Sensitivity and specificity, predictive values, or other accuracy measures	Absolute or relative difference in sensitivity and specificity, predictive values, or other accuracy measures
Relevant for which purposes	Knowing the probability of disease after a test result Finding the most appropriate position for a test in the diagnostic pathway	Estimating the change in accuracy when an alternative test is used Informing decisions on which tests to use*

Footnotes Table 1: Adapted from (3) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). *Additional to the purposes described under 'Accuracy of a single test'. Factors other than comparative accuracy may inform decisions regarding test selection.

3. Development of QUADAS-C

The process of developing QUADAS-C was based on a framework for developing quality assessment tools by Whiting and colleagues (14). A steering group consisting of eight people with a background in diagnostic test evaluation and/or systematic review methodology coordinated all activities. Six

members of the steering group (P.F.W., S.M., J.J.D., M.M.G.L., C.F.D., C.H.) had also been involved in the development of QUADAS-2.

3.1. Delphi study

For achieving consensus on the scope and on which items to include in the tool, we conducted a Delphi study (protocol registered at https://osf.io/tmze9). The study was designed as four rounds of surveys interspersed with feedback of results to all panel members. After each round, the steering group held a teleconference to discuss the results of the previous round and the design of the next round.

We invited international experts in the field of diagnostic test accuracy research to participate in the study, who were identified based on the recommendations of individual steering group members. The 16 experts who accepted our invitation (6 of whom had been involved in the development of QUADAS-2) formed the QUADAS-C advisory group. Together with the 8 members of the steering group, all 24 people participated in the Delphi study as panel members.

Prior to the first Delphi round, the steering group compiled an initial list of items that were considered potentially important for inclusion in the tool. The sources we consulted for identifying potentially important items included: an overview of comparative DTA reviews published in 2017 (3), any risk of bias items associated with comparative DTA studies used in 102 Cochrane DTA review protocols with a comparative question (date of search in Cochrane Library: July 2018), and an article by Wade and colleagues, who described their experience in modifying QUADAS-2 for use in a comparative DTA systematic review (7). Only one meta-epidemiological study provided empirical evidence of potential bias in comparative accuracy research (2). Studies investigating bias in randomized trials of interventions (15) were consulted as indirect evidence for items relating to the randomization process. The initial list of items was finalized during a face-to-face steering group meeting in September 2018 in Edinburgh, UK. This list, containing 16 items, fed into the first Delphi round. Details on the item generating process are available in Supplement 1.

The aims of Delphi rounds 1, 2, and 3 were to collect panel members' opinions regarding the fundamental properties and scope of QUADAS-C, which items to include in the tool, and to generate additional items. Items were included in the tool or excluded from a Delphi round following a pre-defined threshold for consensus (70% agreement). Items not reaching this threshold were re-rated in subsequent rounds with occasional amendments to wording. After round 3, the steering group evaluated all five remaining items for which no consensus had been achieved and decided which items to include, providing justifications to the panel. In round 4, the proposed final list of included items was presented and panel members were invited to comment on the tool. The Delphi study led to

the development of the first draft version of QUADAS-C, which was revised further in a face-to-face consensus meeting. The anonymized results of each Delphi round are available in Supplement 1.

3.2. Consensus meeting

We held a two-day consensus meeting for the QUADAS-C group in August 2019 in Birmingham, UK, which was attended by 16 of 24 members (8 steering group, 8 advisory group members). The main focus of the first day was to resolve remaining issues arising from the Delphi study through small group discussions. Additionally, the group piloted the tool on two comparative DTA studies to identify challenges associated with its practical use. On the second day, the steering group critically reviewed the tool, discussed plans for piloting the tool, and agreed on the terminology to be used in the guidance document. Based on the outcomes of the meeting, the steering group revised QUADAS-C to its publicly pilotable version.

3.3. Pilot study

The last phase of the development was a pilot study to collect users' experiences with and feedback on using QUADAS-C (protocol registered at https://osf.io/agx3z). We recruited participants through various networks including authors of Cochrane Reviews, members of the Cochrane Screening and Diagnostic Tests Methods Group, the GRADE (Grading of Recommendations Assessment, Development and Evaluation) Working Group, our affiliated universities, and Twitter (www.twitter.com). Anyone interested in comparative DTA studies or systematic reviews, including healthcare providers, researchers and students, was invited to pilot QUADAS-C on one of four comparative DTA studies purposely chosen to represent various designs (16–19). We also invited authors of ongoing systematic reviews to try out QUADAS-C in their review. Forty-four people participated in the pilot, of which six piloted the tool in ongoing DTA systematic reviews (one review (20) has been published) or other types of evidence syntheses. Results of the pilot study are available in Supplement 1. While participants generally found the tool to be complete and easy to use, they also highlighted items that were ambiguous or in need of further explanation; this lead us to make changes to item wording and to include brief explanations for each item in the tool. The steering group implemented these last changes and circulated the final version to the advisory group for approval.

3.4. Role of the funding source

Amsterdam UMC (The Netherlands) provided funding for this study. The funding organization had no role in the design, collection, analysis, and interpretation of the data or the decision to approve publication of the finished manuscript.

4. The QUADAS-C tool

The final version of QUADAS-C can be found in Supplement 2 and on www.quadas.org. QUADAS-C is intended to assess the risk of bias of test comparisons undertaken in comparative DTA studies. The tool is designed to be an extension of QUADAS-2, meaning that it should be used together with QUADAS-2, as the risk of bias judgments from QUADAS-2 are required to make risk of bias judgments in QUADAS-C.

QUADAS-C contains 14 signaling questions and 4 risk of bias judgment questions across the same four domains as QUADAS-2: (1) Patient Selection, (2) Index Test, (3) Reference Standard and (4) Flow and Timing (Table 2). In the remainder of this article, we elaborate on the basic principles and structure of QUADAS-C; for a more detailed explanation on how to use the tool, we refer the reader to the Guidance Document in Supplement 3, also to be found on www.quadas.org.

Table 2 provides our proposal on how to use the two tools together. QUADAS-2 is completed multiple times, once for each index test, while QUADAS-C is completed once per comparison. Additional columns can be added in QUADAS-2 for each additional test in the comparison.

Table 2. QUADAS-C together with QUADAS-2.

Domain 1: Patient Selection			
Single test accuracy (QUADAS-2)		**Answers for test A**	**Answers for test B**
Signaling questions	1.1 Was a consecutive or random sample of patients enrolled?	Yes/No/Unclear	Yes/No/Unclear
	1.2 Was a case-control design avoided?	Yes/No/Unclear	Yes/No/Unclear
	1.3 Did the study avoid inappropriate exclusions?	Yes/No/Unclear	Yes/No/Unclear
Risk of bias	1.4 Could the selection of patients have introduced bias?	Low/High/Unclear	Low/High/Unclear
Applicability concerns	1.5 Are there concerns that the included patients do not match the review question?	Low/High/Unclear	Low/High/Unclear
Comparative accuracy (QUADAS-C)		**Answers for the test comparison**	
Signaling questions	C1.1 Was the risk of bias for each index test judged 'low' for this domain?*	Yes/No	
	C1.2 Was a fully paired or randomized design used?	Yes/No/Unclear	
	C1.3 Was the allocation sequence random?†	Yes/No/Unclear/Not applicable	
	C1.4 Was the allocation sequence concealed until patients were enrolled and assigned to index tests?†	Yes/No/Unclear/Not applicable	
Risk of bias	C1.5 Could the selection of patients have introduced bias in the comparison?	Low/High/Unclear	
Domain 2: Index Test			
Single test accuracy (QUADAS-2)		**Answers for test A**	**Answers for test B**
Signaling questions	2.1 Were the index test results interpreted without knowledge of the results of the reference standard?	Yes/No/Unclear	Yes/No/Unclear
	2.2 If a threshold was used, was it prespecified?	Yes/No/Unclear	Yes/No/Unclear
Risk of bias	2.3 Could the conduct or interpretation of the index test have introduced bias?	Low/High/Unclear	Low/High/Unclear
Applicability concerns	2.4 Are there concerns that the index test, its conduct or its interpretation differ from the review question?	Low/High/Unclear	Low/High/Unclear
Comparative accuracy (QUADAS-C)		**Answers for the test comparison**	
Signaling questions	C2.1 Was the risk of bias for each index test judged 'low' for this domain?*	Yes/No	
	C2.2 Were the index test results interpreted without knowledge of the results of the other index test(s)?‡	Yes/No/Unclear/Not applicable	
	C2.3 Is undergoing one index test unlikely to affect the performance of the other index test(s)?‡	Yes/No/Unclear/Not applicable	
	C2.4 Were the index tests conducted and interpreted without advantaging one of the tests?	Yes/No/Unclear	
Risk of bias	C2.5 Could the conduct or interpretation of the index tests have introduced bias in the comparison?	Low/High/Unclear	
Domain 3: Reference Standard			
Single test accuracy (QUADAS-2)		**Answers for test A**	**Answers for test B**
Signaling questions	3.1 Is the reference standard likely to correctly classify the target condition?	Yes/No/Unclear	Yes/No/Unclear
	3.2 Were the reference standard results interpreted without knowledge of the results of the index test?	Yes/No/Unclear	Yes/No/Unclear
Risk of bias	3.3 Could the reference standard, its conduct, or its interpretation have introduced bias?	Low/High/Unclear	Low/High/Unclear
Applicability concerns	3.4 Are there concerns that the target condition as defined by the reference standard does not match the review question?	Low/High/Unclear	Low/High/Unclear
Comparative accuracy (QUADAS-C)		**Answers for the test comparison**	
Signaling questions	C3.1 Was the risk of bias for each index test judged 'low' for this domain?*	Yes/No	
	C3.2 Did the reference standard avoid incorporating any of the index tests?	Yes/No/Unclear	
Risk of bias	C3.3 Could the reference standard, its conduct, or its interpretation have introduced bias in the comparison?	Low/High/Unclear	
Domain 4: Flow and Timing			
Single test accuracy (QUADAS-2)		**Answers for test A**	**Answers for test B**
Signaling questions	4.1 Was there an appropriate interval between index tests and reference standard?	Yes/No/Unclear	Yes/No/Unclear
	4.2 Did all patients receive a reference standard?	Yes/No/Unclear	Yes/No/Unclear
	4.3 Did all patients receive the same reference standard?	Yes/No/Unclear	Yes/No/Unclear
	4.4 Were all patients included in the analysis?	Yes/No/Unclear	Yes/No/Unclear
Risk of bias	4.5 Could the patient flow have introduced bias?	Low/High/Unclear	Low/High/Unclear
Comparative accuracy (QUADAS-C)		**Answers for the test comparison**	
Signaling questions	C4.1 Was the risk of bias for each index test judged 'low' for this domain?*	Yes/No	
	C4.2 Was there an appropriate interval between the index tests?	Yes/No/Unclear	
	C4.3 Was the same reference standard used for all index tests?	Yes/No/Unclear	
	C4.4 Are the proportions and reasons for missing data similar across index tests?	Yes/No/Unclear	
Risk of bias	C4.5 Could the patient flow have introduced bias in the comparison?	Low/High/Unclear	

Footnote to table 2:
* Refers back to the QUADAS-2 risk of bias judgments (questions 1.4, 2.3, 3.3, or 4.5)
† Only applicable to randomized designs.
‡ Only applicable if patients received multiple index tests (fully or partially paired designs)

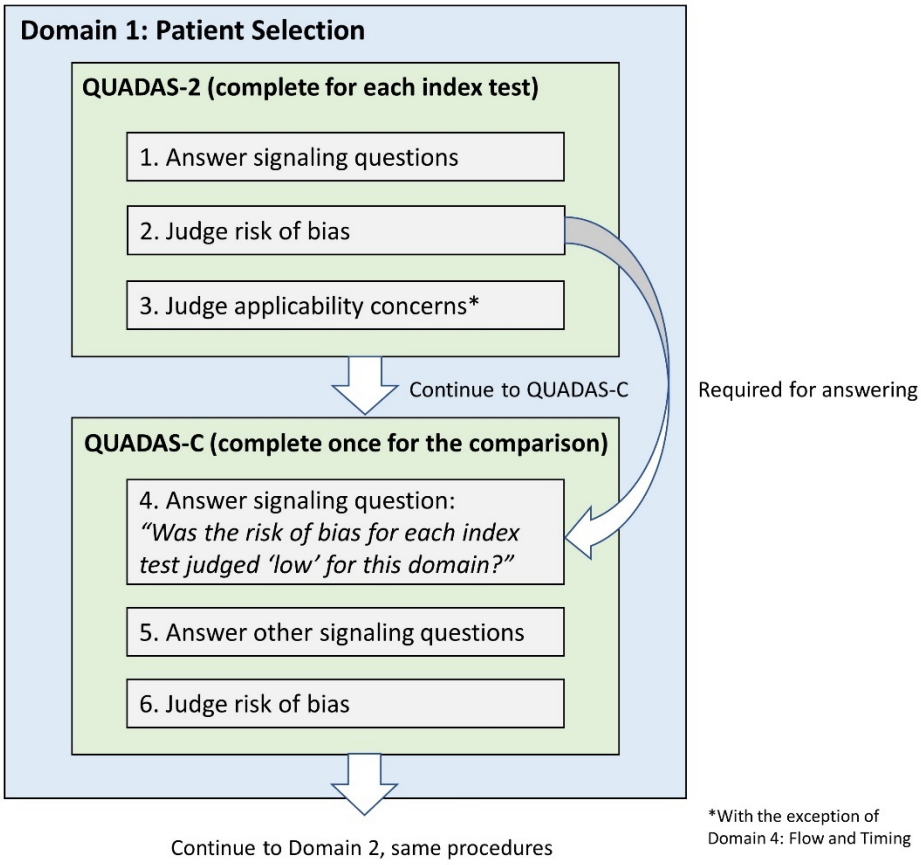
4.1. Scope of QUADAS-C

QUADAS-C was designed primarily with assessment of fully paired and randomized comparative DTA studies in mind. Taken together, these comprise the majority of comparative DTA study designs in systematic reviews (10). While QUADAS-C could be used to assess other comparative DTA designs, the tool will need to be tailored to the specific design being assessed, for example by including new signaling questions and removing irrelevant ones. Particularly in unpaired or partially paired studies without randomization, the issue of confounding will need to be addressed in more detail. QUADAS-C is not designed to assess risk of bias in test comparisons made between studies (also called *between-study* or *indirect* comparisons).

Some comparative DTA studies may include a comparison between one or more testing strategies (i.e. combinations of tests), to assess whether one testing strategy is more accurate than another test or testing strategy. QUADAS-C can be used to assess these comparisons as well, though users will need to define the comparison clearly and careful tailoring of the tool may be required.

In contrast to QUADAS-2, QUADAS-C does not have questions on concerns regarding applicability. Users can nevertheless arrive at a judgment regarding applicability of the test comparison by choosing the highest concern (i.e. the worst) applicability judgment for an index test in QUADAS-2. For example, an item in the Index Test domain of QUADAS-2 is: '*Is there concern that the index test, its conduct, or interpretation differ from the review question?*'. If the answer for test A is 'low concern' and the answer for test B is 'high concern', it is clear that there is high concern regarding the applicability of the comparison between A and B. The assessment of applicability, although not part of QUADAS-C, is no less important than the assessment of risk of bias. Therefore, we strongly recommend users to also consider and describe the applicability of test comparisons, based on applicability judgments from QUADAS-2.

Figure 1. Process of using QUADAS-2 and QUADAS-C together.



4.2. How is QUADAS-C used together with QUADAS-2?

Figure 1 shows schematically how QUADAS-2 and QUADAS-C are completed together when assessing a comparative DTA study. First, starting with the Patient Selection domain, QUADAS-2 is completed for each index test separately. Assuming a comparison of two tests, this will lead to risk of bias and applicability judgments for test A, followed by risk of bias and applicability judgments for test B.

Next, still within the Patient Selection domain, QUADAS-C is completed once for the comparison between tests A and B. The first signaling question of each domain in QUADAS-C, *"Was the risk of bias for each index test judged 'low' for this domain?"*, makes use of the risk of bias judgments in QUADAS-2: if both risk of bias judgments for test A and test B were 'low', this question is answered 'yes', implying a low risk of bias for the comparison. By subsequently answering other QUADAS-C signaling questions in this domain, users can reach a risk of bias judgment for the comparison. The same procedure is repeated for subsequent domains (Index Test, Reference Standard, Flow and Timing).

By having the signaling question *"Was the risk of bias for each index test judged 'low' for this domain?"* in each domain, QUADAS-C requires a low risk of bias judgment for each index test in QUADAS-2 for a low risk of bias judgment in QUADAS-C. When the risk of bias is 'high' for one or both index tests in QUADAS-2, potential for bias in the comparison exists. Although it may be possible that the direction and magnitude of bias affecting each index test may cancel each other out, such predictions are difficult to make.

4.3. Assessing risk of bias with QUADAS-C

We recommend users to complete QUADAS-C in four phases, similar to the process for QUADAS-2: 1) clearly state the review question, 2) tailor the tool to each review and develop review-specific guidance, 3) review the study flow diagram or construct one if none is reported, and 4) judge risk of bias. The Guidance Document in Supplement 3 provides details of each phase. Whenever a study includes multiple comparisons of interest, QUADAS-C needs to be completed for each one of those comparisons, since a risk of bias judgment in QUADAS-C is specific to a particular test comparison.

4.3.1. Information to support the judgment of risk of bias

When judging risk of bias, users should record all the information used to reach the judgment for reasons of transparency and reproducibility. For this purpose, QUADAS-C contains free text fields for recording 1) the comparative study design (users can choose from a set of prespecified designs or describe the design) and 2) information relevant to the validity of the comparison. The latter should be recorded for each of the four domains: for instance, how participants were allocated to index tests (Patient Selection domain), and whether there were any differences in the reasons for missing data between index tests (Flow and Timing domain).

4.3.2. Answering signaling questions

Each signaling question in QUADAS-C can be answered 'yes', 'no', or 'unclear', where 'yes' indicates low risk of bias. A 'no' answer implies that potential for bias exists, but it does not automatically lead to a high risk of bias judgment for that domain; instead, users need to consider the likelihood and importance of the bias (see also section 4.3.3). The options 'yes' and 'no' should also be used when the user's assessment is 'probably yes' or 'probably no', respectively. The option 'unclear' is only appropriate if there is insufficient information to answer either 'yes' or 'no'. Detailed explanations with examples for answering each signaling question are provided in the Guidance Document (Supplement 3).

4.3.3. Judging the risk of bias for each domain

The answers to signaling questions will help the user to arrive at a risk of bias judgment for each domain, which can be 'low', 'high', or 'unclear'. A 'yes' answer to all signaling questions within a

domain should typically lead to a low risk of bias judgment. A 'no' answer to a single signaling question may lead to a high risk of bias judgment if the associated bias is of such concern that the entire domain is deemed problematic; this is indeed often a judgment call on the users' part. Users may judge risk of bias as 'unclear' if there is insufficient information to judge as either low or high risk.

4.3.4. Judging the overall risk of bias across all domains

While not formally part of QUADAS-C, users may find it helpful to produce an overall risk of bias judgment across all domains for each study. An example would be to judge 'low overall risk of bias' if all domains were at low risk of bias, and to judge 'high' or 'unclear overall risk of bias' if one or more domains were at high or unclear risk of bias, respectively.

4.4. Incorporating QUADAS-C assessments in comparative DTA systematic reviews

4.4.1. Narrative and visual summaries of risk of bias judgments

Users of QUADAS-C are strongly encouraged to provide a narrative and/or visual summary of their risk of bias judgments across studies. Table 3 is an example of presenting QUADAS-2 and QUADAS-C results together. If the comparison is between two index tests, the combined use of QUADAS-2 and QUADAS-C will result in 1) judgments for the accuracy of test A, 2) judgments for the accuracy of test B, and 3) judgments for the comparison between A and B. If the review question only concerns comparative accuracy, users may decide to display only QUADAS-C results. The Guidance Document (Supplement 3) contains additional suggestions on how to present results.

4.4.2. Using risk of bias judgments to inform the analysis, conclusions, and the certainty of evidence

Risk of bias judgments can be used to investigate between-study heterogeneity (either by subgroup analysis or meta-regression) or to explore the impact of excluding particular studies from meta-analyses (21). Such analyses can be done using risk of bias judgments for a particular domain or overall risk of bias judgments across domains. For example, users may assess whether studies at high risk of bias show different relative accuracy compared to studies at low risk of bias. Users may decide to exclude studies at high risk of bias from the primary analysis or as a sensitivity analysis. Ideally, QUADAS-C results should also be incorporated in the conclusions of systematic reviews (22). Risk of bias judgments can further inform assessments of the certainty, quality, or strength of the overall body of evidence (23).

Table 3. Suggestion on how to present QUADAS-2 and QUADAS-C results together.

Study	Test	Risk of bias (QUADAS-2)				Applicability concerns (QUADAS-2)			Risk of bias (QUADAS-C)			
		P	I	R	FT	P	I	R	P	I	R	FT
Author, year	A	✓	✗	✓	✓	✓	✓	✓	✗	✗	✓	✓
	B	✓	✓	✓	✓	✓	✓	✗				
Author, year	A	?	✓	✓	✗	✓	?	✓	?	✗	✓	✗
	B	?	✓	✓	✗	✓	✓	✓				
Author, year	A	✓	✓	✓	✓	?	✓	✓	✓	?	?	✓
	B	✓	?	✓	✓	?	✓	✓				

Footnote to table 3:
P = Patient Selection; I = Index Test; R = Reference Standard; FT = Flow and Timing.
✓ indicates low risk; ✗ indicates high risk; ? indicates unclear risk.
The current table may be simplified if the QUADAS-2 judgments for P, R, and FT are the same for each index test. See Supplement 3 for this and other examples on how to present results. Templates for tabular and graphical presentations are available at www.quadas.org.

5. Discussion

Decisions regarding the selection of diagnostic tests for clinical practice may benefit from trustworthy evidence on the relative accuracy of alternative tests. While comparative DTA studies can provide valid evidence on relative test performance, it is essential that such studies are critically evaluated for any shortcomings in their design, conduct, and analysis that may bias their results.

We developed QUADAS-C through a rigorous process of iterative feedback, consensus, and user testing. QUADAS-C is explicitly developed with the structure and design of QUADAS-2 in mind, so that users who have experience with QUADAS-2 may find the extension straightforward to use. We acknowledge that the items in QUADAS-C are mainly based on consensus and theoretical considerations; empirical confirmation of bias is still limited. Like many quality assessment tools, we expect that QUADAS-C will need updating as knowledge regarding biases in comparative DTA studies evolve over time.

QUADAS-C has been designed as a generic tool for comparing all types of diagnostic tests. As unique methodological considerations may apply to specific diagnostic tests, users are invited to tailor the tool to the individual systematic review by adding, omitting, or modifying signaling questions. For example, PROBAST (Prediction model Risk Of Bias ASsessment Tool) (24) provides more specific signaling questions for multivariable models which users could consider when tailoring.

It should be noted that QUADAS-C is not appropriate for assessing the risk of bias in studies that evaluate the effectiveness of test-treatment strategies on people-important outcomes, such as morbidity and mortality. For those studies, users should use tools matching the type of study, such as the revised Cochrane risk of bias tool for randomized trials (25) and ROBINS-I (Risk Of Bias In Non-randomised Studies - of Interventions) for nonrandomized studies of interventions (26).

As observed during the Delphi rounds and the pilot study, the use of QUADAS-C is not without challenges. As the tool is used together with QUADAS-2, users (especially those who are unfamiliar with QUADAS-2) may find the combined number of signaling questions quite large. Furthermore, assessing the risk of bias in test comparisons with three or more tests, while possible, may be challenging. QUADAS-C was designed with fully paired and randomized studies in mind, and its use for assessing nonrandomized and other 'creative' designs will require additional tailoring of the tool. The development of a web-based tool, which is currently planned, may resolve some of the issues raised by users, such as automated completion of conditional signaling questions, optional display of explanations to signaling questions, and automated construction of exportable risk of bias tables and graphs that combine QUADAS-2 and QUADAS-C results.

We hope that QUADAS-C will help review authors to systematically perform risk of bias assessments and identify high-quality studies in comparative DTA systematic reviews, help primary study investigators avoid potential biases in the design and conduct of their study and, more generally, increase awareness of the importance of methodological rigor among those involved in comparative accuracy research. It may also raise awareness that comparing test accuracy using estimates obtained from noncomparative studies, a common practice in systematic reviews (3), is intrinsically at risk of generating biased results, reinforcing the need for well-designed comparative DTA studies to inform decision-making regarding preferred tests.

6. Acknowledgements

We thank all participants in our pilot study for providing us with insightful feedback on improving the draft version of QUADAS-C. In addition, we would like to thank Jenny Lee, MSc (University of Amsterdam) for providing valuable feedback and suggestions on this manuscript. Y.T. is supported by a National Institute for Health Research (NIHR) Postdoctoral Fellowship. Y.T., J.D., and C.D. are supported by the NIHR Birmingham Biomedical Research Centre. This article presents independent research supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

7. References

1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. Clin Chem. 2003 Jan 1;49(1):7–18.
2. Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical Evidence of the Importance of Comparative Studies of Diagnostic Test Accuracy. Ann Intern Med. 2013 Apr 2;158(7):544.
3. Yang B, Vali Y, Dehmoobad Sharifabadi A, Harris IM, Beese S, Davenport C, et al. Risk of bias assessment of test comparisons was uncommon in comparative accuracy systematic reviews: an overview of reviews. J Clin Epidemiol. 2020 Nov 1;127:167–74.
4. Leeflang MMG, Reitsma JB. Systematic reviews and meta-analyses addressing comparative test accuracy questions. Diagnostic Progn Res. 2018 Dec 10;2(1):17.
5. Dehmoobad Sharifabadi A, Leeflang M, Treanor L, Kraaijpoel N, Salameh J-P, Alabousi M, et al. Comparative reviews of diagnostic test accuracy in imaging research: evaluation of current practices. Eur Radiol. 2019 Oct;29(10):5386–94.
6. Whiting PF, Rutjes AWS, 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 Oct 18;155(8):529.
7. 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 Sep;4(3):280–6.
8. Donovan J, Thu DDA, Phu NH, Dung VTM, Quang TP, Nghia HDT, et al. Xpert MTB/RIF Ultra versus Xpert MTB/RIF for the diagnosis of tuberculous meningitis: a prospective, randomised, diagnostic accuracy study. Lancet Infect Dis. 2020;20(3):299–307.
9. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006 May 6;332(7549):1089–92.
10. Yang B, Olsen M, Vali Y, Langendam MW, Takwoingi Y, Hyde CJ, et al. Study designs for comparative diagnostic test accuracy: A methodological review and classification scheme. J Clin Epidemiol. 2021 Oct;138:128–38.
11. Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic tests for replacement, add-on, and triage. J Clin Epidemiol. 2010;63(8):883–91.
12. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352:3–7.
13. Takwoingi Y. Meta-analytic approaches for summarising and comparing the accuracy of medical tests. University of Birmingham Research Archive. 2016.
14. Whiting P, Wolff R, Mallett S, Simera I, Savović J. A proposed framework for developing

quality assessment tools. Syst Rev. 2017;6(1):1–9.
15. Page MJ, Higgins JPT, Clayton G, Sterne JAC, Hróbjartsson A, Savović J. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. PLoS One. 2016;11(7):e0159267.
16. Sivit CJ, Applegate KE, Stallion A, Dudgeon DL, Salvator A, Schluchter M, et al. Imaging Evaluation of Suspected Appendicitis in a Pediatric Population. Am J Roentgenol. 2000;175(4):977–80.
17. Kaiser S, Frenckner B, Jorulf HK. Suspected Appendicitis in Children: US and CT— A Prospective Randomized Study. Radiology. 2002;223:633–8.
18. Fazal MA, Khan I, Thomas C. Ultrasonography and Magnetic resonance imaging in the diagnosis of Morton's neuroma. J Am Podiatr Med Assoc. 2012;102(3):184–6.
19. Norton ME, Jacobsson B, Swamy GK, Laurent LC, Ranzini AC, Brar H, et al. Cell-free DNA Analysis for Noninvasive Examination of Trisomy. N Engl J Med. 2015;372(17):1589–97.
20. Zifodya JS, Kreniske JS, Schiller I, Kohli M, Dendukuri N, Schumacher SG, et al. Xpert Ultra versus Xpert MTB/RIF for pulmonary tuberculosis and rifampicin resistance in adults with presumptive pulmonary tuberculosis. Cochrane Database Syst Rev. 2021 Feb 22;2021(2).
21. Boutron I, Page MJ, Higgins JP, Altman DG, Lundh A, Hróbjartsson A. Chapter 7: Considering bias and conflicts of interest among the included studies [Internet]. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). Cochrane; 2020 [cited 2020 Oct 30]. Available from: www.training.cochrane.org/handbook
22. Ochodo EA, Van Enst WA, Naaktgeboren CA, De Groot JAH, Hooft L, Moons KGM, et al. Incorporating quality assessments of primary studies in the conclusions of diagnostic accuracy reviews: A cross-sectional study. BMC Med Res Methodol. 2014;14(1):1–8.
23. Yang B, Mustafa RA, Bossuyt PM, Brozek J, Hultcrantz M, Leeflang MMG, et al. GRADE Guidance: 31. Assessing the certainty across a body of evidence for comparative test accuracy. J Clin Epidemiol. 2021 Aug 14;136:146–56.
24. Wolff RF, Moons KGM, 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. Ann Intern Med. 2019;170(1):51–8.
25. Sterne JAC, 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;366:1–8.
26. 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;355:4–10.

8. Author contributions

B.Y.: Conceptualization, Project Administration, Methodology, Data Collection, Formal Analysis, Writing – Original Draft, Writing – Review & Editing. **S.M.:** Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing. **Y.T.:** Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing. **C.F.D.:** Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing. **C.J.H.:** Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing. **P.F.W.:** Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing. **J.J.D.:** Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing. **M.M.G.L.:** Conceptualization, Project Administration, Methodology, Formal Analysis, Writing – Review & Editing, Supervision.