

## 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](https://doi.org/10.7326/M21-2234)

License:  
None: All rights reserved

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>

[Link to publication on Research at Birmingham portal](#)

### Publisher Rights Statement:

This is an accepted manuscript version of an article first published in *Annals of Internal Medicine*. The final version of record is available at <https://doi.org/10.7326/M21-2234>

### General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

### Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact [UBIRA@lists.bham.ac.uk](mailto:UBIRA@lists.bham.ac.uk) providing details and we will remove access to the work immediately and investigate.

1

2                   **QUADAS-C: a tool for assessing risk of bias in**

3                   **comparative diagnostic accuracy studies**

4

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

8 <sup>a</sup> Department of Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, 1105AZ,  
9 Amsterdam, The Netherlands

10 <sup>b</sup> UCL Centre for Medical Imaging, University College London, London, W1W 7TY, UK

11 <sup>c</sup> Test Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, Edgbaston,  
12 Birmingham, B15 2TT, UK

13 <sup>d</sup> NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University  
14 of Birmingham, Birmingham, UK

15 <sup>e</sup> Exeter Test Group, Institute of Health Research, College of Medicine and Health, University of Exeter, Exeter, UK

16 <sup>f</sup> Population Health Sciences, Bristol Medical School, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8  
17 2PS, UK

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

19 **Corresponding author:**

20 Bada Yang

21 Department of Epidemiology and Data Science, Room J1b-210

22 Amsterdam UMC, Location AMC

23 Meibergdreef 9, 1105AZ Amsterdam, The Netherlands

24 Tel: +31(0)20 5666948

25 Email: [b.d.yang@outlook.com](mailto:b.d.yang@outlook.com)

26

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

53 **Abstract**

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

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

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

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

73

74 Abstract word count: 247

75 Manuscript word count: 3631

76

77 Figures: 1

78 Tables: 3

79 References: 26

80 Running title: QUADAS-C

81 Keywords: Diagnostic accuracy; Bias; Test comparison; Methodology; Systematic review

82 Supplement 1: Information contributing to the development of QUADAS-C

83 Supplement 2: The QUADAS-C tool

84 Supplement 3: Guidance on how to use QUADAS-C

## 85 **1. Introduction**

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

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

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

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

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

112

## 113 **2. Comparative accuracy questions**

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

118 meningitis compared to Xpert® MTB/RIF (8). For a valid comparison, participants receiving index  
 119 test A should be exchangeable with participants receiving index test B. This can be accomplished by  
 120 each participant undergoing all index tests (often referred to as a fully paired or within-subject  
 121 design), or approximated by randomly allocating participants to index tests (randomized design)  
 122 (2,9,10). Comparative accuracy results can be expressed as absolute or relative differences in  
 123 sensitivity and specificity, predictive values, area under the curve, or other measures of accuracy  
 124 including decision analytic measures such as net benefit (11,12). Knowledge about comparative  
 125 accuracy is important for the selection and recommendation of a test from a number of alternative,  
 126 competing tests, especially when studies evaluating the effectiveness of test-treatment strategies on  
 127 patient-important outcomes are absent (13). A key characteristic of comparative accuracy questions is  
 128 that none of the tests being compared is the reference standard. Rather, the reference standard is a  
 129 means to verify whether participants have the target condition or not.

130

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

	<b>Accuracy of a single test</b>	<b>Comparative accuracy</b>
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*

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

135

136

### 137 **3. Development of QUADAS-C**

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

141 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  
142 in the development of QUADAS-2.

### 143 ***3.1. Delphi study***

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

149 We invited international experts in the field of diagnostic test accuracy research to participate in the  
150 study, who were identified based on the recommendations of individual steering group members. The  
151 16 experts who accepted our invitation (6 of whom had been involved in the development of  
152 QUADAS-2) formed the QUADAS-C advisory group. Together with the 8 members of the steering  
153 group, all 24 people participated in the Delphi study as panel members.

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

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

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

### 176 **3.2. Consensus meeting**

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

### 185 **3.3. Pilot study**

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

201



202 **3.4. Role of the funding source**

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

#### 206 **4. The QUADAS-C tool**

207 The final version of QUADAS-C can be found in [Supplement 2](#) and on [www.quadas.org](http://www.quadas.org). QUADAS-  
208 C is intended to assess the risk of bias of test comparisons undertaken in comparative DTA studies.  
209 The tool is designed to be an extension of QUADAS-2, meaning that it should be used together with  
210 QUADAS-2, as the risk of bias judgments from QUADAS-2 are required to make risk of bias  
211 judgments in QUADAS-C.

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

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

220

221

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

<b>Domain 1: Patient Selection</b>			
<b>Single test accuracy (QUADAS-2)</b>		<b>Answers for test A</b>	<b>Answers for test B</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
<b>Comparative accuracy (QUADAS-C)</b>		<b>Answers for the test comparison</b>	
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	
<b>Domain 2: Index Test</b>			
<b>Single test accuracy (QUADAS-2)</b>		<b>Answers for test A</b>	<b>Answers for test B</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
<b>Comparative accuracy (QUADAS-C)</b>		<b>Answers for the test comparison</b>	
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	
<b>Domain 3: Reference Standard</b>			
<b>Single test accuracy (QUADAS-2)</b>		<b>Answers for test A</b>	<b>Answers for test B</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
<b>Comparative accuracy (QUADAS-C)</b>		<b>Answers for the test comparison</b>	
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	
<b>Domain 4: Flow and Timing</b>			
<b>Single test accuracy (QUADAS-2)</b>		<b>Answers for test A</b>	<b>Answers for test B</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
<b>Comparative accuracy (QUADAS-C)</b>		<b>Answers for the test comparison</b>	
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	

223 Footnote to table 2:

224 \* Refers back to the QUADAS-2 risk of bias judgments (questions 1.4, 2.3, 3.3, or 4.5)

225 † Only applicable to randomized designs.

226 ‡ Only applicable if patients received multiple index tests (fully or partially paired designs)

227

228 **4.1. Scope of QUADAS-C**

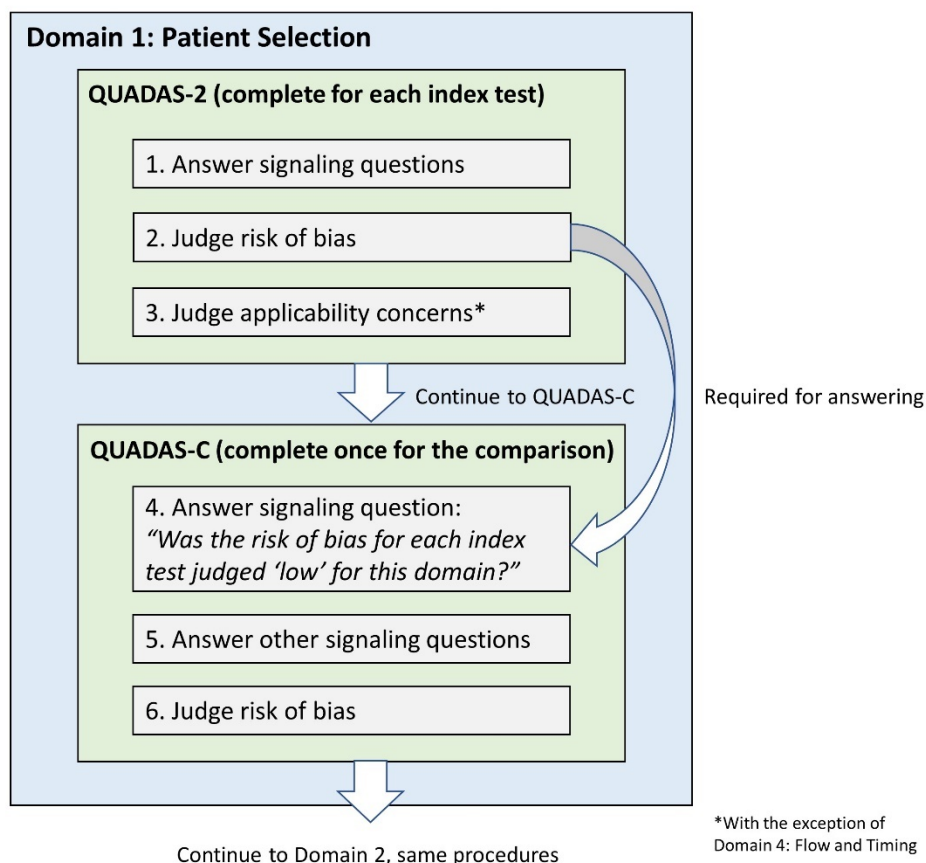
229 QUADAS-C was designed primarily with assessment of fully paired and randomized comparative  
230 DTA studies in mind. Taken together, these comprise the majority of comparative DTA study designs  
231 in systematic reviews (10). While QUADAS-C could be used to assess other comparative DTA  
232 designs,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

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

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

251

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



254 **4.2. How is QUADAS-C used together with QUADAS-2?**

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

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

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

### 274 **4.3. Assessing risk of bias with QUADAS-C**

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

#### 281 4.3.1. Information to support the judgment of risk of bias

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

#### 289 4.3.2. Answering signaling questions

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

#### 298 4.3.3. Judging the risk of bias for each domain

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

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

#### 306 4.3.4. Judging the overall risk of bias across all domains

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

### 311 ***4.4. Incorporating QUADAS-C assessments in comparative DTA systematic reviews***

#### 312 4.4.1. Narrative and visual summaries of risk of bias judgments

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

#### 320 4.4.2. Using risk of bias judgments to inform the analysis, conclusions, and the certainty of evidence

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

330

331 **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	✓	?	✓	✓	?	✓	✓				

332 Footnote to table 3:

333 P = Patient Selection; I = Index Test; R = Reference Standard; FT = Flow and Timing.

334 ✓ indicates low risk; ✗ indicates high risk; ? indicates unclear risk.

335 The current table may be simplified if the QUADAS-2 judgments for P, R, and FT are the same for each index

336 test. See [Supplement 3](#) for this and other examples on how to present results. Templates for tabular and

337 graphical presentations are available at [www.quadas.org](http://www.quadas.org).

338

## 339 5. Discussion

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

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

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

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



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

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

379

## 380 **6. Acknowledgements**

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

390

391

## 392 7. References

- 393 1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD  
394 Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. *Clin*  
395 *Chem.* 2003 Jan 1;49(1):7–18.
- 396 2. Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical Evidence of the Importance of  
397 Comparative Studies of Diagnostic Test Accuracy. *Ann Intern Med.* 2013 Apr 2;158(7):544.
- 398 3. Yang B, Vali Y, Dehmoobad Sharifabadi A, Harris IM, Beese S, Davenport C, et al. Risk of  
399 bias assessment of test comparisons was uncommon in comparative accuracy systematic  
400 reviews: an overview of reviews. *J Clin Epidemiol.* 2020 Nov 1;127:167–74.
- 401 4. Leeflang MMG, Reitsma JB. Systematic reviews and meta-analyses addressing comparative  
402 test accuracy questions. *Diagnostic Progn Res.* 2018 Dec 10;2(1):17.
- 403 5. Dehmoobad Sharifabadi A, Leeflang M, Treanor L, Kraaijpoel N, Salameh J-P, Alabousi M, et  
404 al. Comparative reviews of diagnostic test accuracy in imaging research: evaluation of current  
405 practices. *Eur Radiol.* 2019 Oct;29(10):5386–94.
- 406 6. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2:  
407 A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med.*  
408 2011 Oct 18;155(8):529.
- 409 7. Wade R, Corbett M, Eastwood A. Quality assessment of comparative diagnostic accuracy  
410 studies: our experience using a modified version of the QUADAS-2 tool. *Res Synth Methods.*  
411 2013 Sep;4(3):280–6.
- 412 8. Donovan J, Thu DDA, Phu NH, Dung VTM, Quang TP, Nghia HDT, et al. Xpert MTB/RIF  
413 Ultra versus Xpert MTB/RIF for the diagnosis of tuberculous meningitis: a prospective,  
414 randomised, diagnostic accuracy study. *Lancet Infect Dis.* 2020;20(3):299–307.
- 415 9. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against  
416 existing diagnostic pathways. *BMJ.* 2006 May 6;332(7549):1089–92.
- 417 10. Yang B, Olsen M, Vali Y, Langendam MW, Takwoingi Y, Hyde CJ, et al. Study designs for  
418 comparative diagnostic test accuracy: A methodological review and classification scheme. *J*  
419 *Clin Epidemiol.* 2021 Oct;138:128–38.
- 420 11. Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to  
421 assess diagnostic tests for replacement, add-on, and triage. *J Clin Epidemiol.* 2010;63(8):883–  
422 91.
- 423 12. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of  
424 prediction models, molecular markers, and diagnostic tests. *BMJ.* 2016;352:3–7.
- 425 13. Takwoingi Y. Meta-analytic approaches for summarising and comparing the accuracy of  
426 medical tests. University of Birmingham Research Archive. 2016.
- 427 14. Whiting P, Wolff R, Mallett S, Simera I, Savović J. A proposed framework for developing

- 428 quality assessment tools. *Syst Rev.* 2017;6(1):1–9.
- 429 15. Page MJ, Higgins JPT, Clayton G, Sterne JAC, Hróbjartsson A, Savović J. Empirical Evidence  
430 of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological  
431 Studies. *PLoS One.* 2016;11(7):e0159267.
- 432 16. Sivit CJ, Applegate KE, Stallion A, Dudgeon DL, Salvator A, Schluchter M, et al. Imaging  
433 Evaluation of Suspected Appendicitis in a Pediatric Population. *Am J Roentgenol.*  
434 2000;175(4):977–80.
- 435 17. Kaiser S, Frenckner B, Jorulf HK. Suspected Appendicitis in Children: US and CT— A  
436 Prospective Randomized Study. *Radiology.* 2002;223:633–8.
- 437 18. Fazal MA, Khan I, Thomas C. Ultrasonography and Magnetic resonance imaging in the  
438 diagnosis of Morton’s neuroma. *J Am Podiatr Med Assoc.* 2012;102(3):184–6.
- 439 19. Norton ME, Jacobsson B, Swamy GK, Laurent LC, Ranzini AC, Brar H, et al. Cell-free DNA  
440 Analysis for Noninvasive Examination of Trisomy. *N Engl J Med.* 2015;372(17):1589–97.
- 441 20. Zifodya JS, Kreniske JS, Schiller I, Kohli M, Dendukuri N, Schumacher SG, et al. Xpert Ultra  
442 versus Xpert MTB/RIF for pulmonary tuberculosis and rifampicin resistance in adults with  
443 presumptive pulmonary tuberculosis. *Cochrane Database Syst Rev.* 2021 Feb 22;2021(2).
- 444 21. Boutron I, Page MJ, Higgins JP, Altman DG, Lundh A, Hróbjartsson A. Chapter 7:  
445 Considering bias and conflicts of interest among the included studies [Internet]. Higgins J,  
446 Thomas J, Chandler J, Cumpston M, Li T, Page M, et al., editors. *Cochrane Handbook for*  
447 *Systematic Reviews of Interventions* version 6.1 (updated September 2020). Cochrane; 2020  
448 [cited 2020 Oct 30]. Available from: [www.training.cochrane.org/handbook](http://www.training.cochrane.org/handbook)
- 449 22. Ochodo EA, Van Enst WA, Naaktgeboren CA, De Groot JAH, Hooft L, Moons KGM, et al.  
450 Incorporating quality assessments of primary studies in the conclusions of diagnostic accuracy  
451 reviews: A cross-sectional study. *BMC Med Res Methodol.* 2014;14(1):1–8.
- 452 23. Yang B, Mustafa RA, Bossuyt PM, Brozek J, Hultcrantz M, Leeftang MMG, et al. GRADE  
453 Guidance: 31. Assessing the certainty across a body of evidence for comparative test accuracy.  
454 *J Clin Epidemiol.* 2021 Aug 14;136:146–56.
- 455 24. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST:  
456 A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med.*  
457 2019;170(1):51–8.
- 458 25. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: A revised  
459 tool for assessing risk of bias in randomised trials. *BMJ.* 2019;366:1–8.
- 460 26. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-  
461 I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ.*  
462 2016;355:4–10.
- 463

464 **8. Author contributions**

465 **B.Y.:** Conceptualization, Project Administration, Methodology, Data Collection, Formal Analysis,  
466 Writing – Original Draft, Writing – Review & Editing. **S.M.:** Conceptualization, Methodology,  
467 Formal Analysis, Writing – Review & Editing. **Y.T.:** Conceptualization, Methodology, Formal  
468 Analysis, Writing – Review & Editing. **C.F.D.:** Conceptualization, Methodology, Formal Analysis,  
469 Writing – Review & Editing. **C.J.H.:** Conceptualization, Methodology, Formal Analysis, Writing –  
470 Review & Editing. **P.F.W.:** Conceptualization, Methodology, Formal Analysis, Writing – Review &  
471 Editing. **J.J.D.:** Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing.  
472 **M.M.G.L.:** Conceptualization, Project Administration, Methodology, Formal Analysis, Writing –  
473 Review & Editing, Supervision.