Guidance on how to use QUADAS-C

A tool for assessing risk of bias in comparative diagnostic accuracy studies

Developed by Bada Yang, Sue Mallett, Yemisi Takwoingi, Clare F. Davenport, Christopher J. Hyde, Penny F. Whiting, Jonathan J. Deeks, Mariska M.G. Leeflang, and the QUADAS-C Group*

**QUADAS-C Group: Patrick M.M. Bossuyt, Miriam Brazzelli, Clare F. Davenport, Jonathan J. Deeks, Jacqueline Dinnes, Kurinchi S. Gurusamy, Hayley E. Jones, Christopher J. Hyde, Stefan Lange, Miranda W. Langendam, Mariska M.G. Leeflang, Petra Macaskill, Sue Mallett, Matthew D.F. McInnes, Johannes B. Reitsma, Anne W.S. Rutjes, Alison Sinclair, Yemisi Takwoingi, Henrica C.W. de Vet, Gianni Virgilli, Ros Wade, Marie E. Westwood, Penny F. Whiting, and Bada Yang*

Correspondence: Bada Yang (b.d.yang@outlook.com)

Contents

Please cite this document as: *Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF, Deeks JJ, Leeflang MMG, and the QUADAS-C Group. Guidance on how to use QUADAS-C. 2021 Sept 04. Available from:* *https://osf.io/hq8mf/files/*

1. Background

1.1. Context

Systematic reviews of diagnostic test accuracy (DTA) studies often include studies assessing the performance of a single test against the reference standard. The risk of bias and concerns regarding applicability of these single index test DTA studies are assessed using the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) tool (1).

When selecting and recommending which diagnostic tests to use in clinical practice, it may sometimes be more important to know how the accuracy of one test compares to that of other tests. Information regarding comparative accuracy is best derived from studies that compare two or more index tests in a single study, also called comparative DTA studies. QUADAS-2 was not specifically developed to assess the risk of bias in comparative DTA studies, and additional sources of bias regarding comparisons are not addressed in the tool (Table 1).

We have developed QUADAS-C (C stands for comparative) to assess risk of bias in comparative DTA studies. QUADAS-C is designed to be an extension (i.e. a set of additional questions) to QUADAS-2. This means QUADAS-C should be used together with QUADAS-2, resulting in separate risk of bias judgments for single test accuracy estimates and for comparative accuracy estimates. QUADAS-C assesses the risk of bias in the same four domains as QUADAS-2: Patient Selection, Index Test, Reference Standard, and Flow and Timing.

1.2. Scope of QUADAS-C

There are several comparative DTA study designs; these are described in Section 1.3. QUADAS-C was designed primarily with assessment of 'fully paired' and 'randomized' designs in mind. Taken together, these comprise the majority of comparative DTA study designs included in systematic reviews (2). 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.

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.

QUADAS-C is designed to assess risk of bias, but not to identify concerns regarding applicability. We believe that concerns regarding applicability related to the comparison can be inferred from the judgments for each individual test, as assessed using QUADAS-2. The view of the QUADAS-C group is that, if at least one of the index tests has a 'high concern' regarding applicability, the comparison would have a 'high concern' as well. As the assessment of applicability is no less important than the assessment of risk of bias, we recommend that review authors always consider and describe the concerns regarding applicability of test comparisons in their reviews.

QUADAS-C is not designed for an assessment of comparisons made *between* separate studies (as opposed to comparisons of tests *within* a study), with each study including only one of the index tests being compared. Such between-study comparisons (also called indirect comparisons) are commonly undertaken in systematic reviews (3), but the bias in these comparisons does not result from study deficiencies; rather it is a feature of the indirect nature of a comparison in a systematic review. Therefore, between-study comparisons are outside the scope of QUADAS-C. The GRADE Working Group recently developed guidance on how to rate the certainty of evidence in between-study comparisons of test accuracy (4).

Table 1. Differences in sources of bias between noncomparative (single index test) diagnostic test accuracy studies and comparative (multiple index test) diagnostic test accuracy studies.

	Noncomparative diagnostic test accuracy study	**Comparative diagnostic test accuracy study**
Health-related question	How accurately can an index test classify individuals with and without 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 the index test and the reference standard	A study in which participants are consecutively or randomly sampled and: (I) each participant undergoes all index tests and the reference standard (fully paired or within-subject design) or (II) participants are randomly allocated to one of the index tests and all participants receive the reference standard (randomized design)
Examples of risk of bias*	The participant spectrum of those who have the target condition only includes advanced disease	The participant spectrum differs between those who undergo index test A and those who undergo index test B
	The index test is interpreted with knowledge of the results of the reference standard	Index test A is interpreted with knowledge of the results of index test B
	The reference standard does not correctly classify the target condition	Results of index test A and B are verified against different reference standards
	There is an inappropriate time interval between the index test and the reference standard	There is an inappropriate time interval between index test A and index test B
	Not all participants enrolled in the study are included in the analysis	There are different proportions of missing data for index test A and index test B

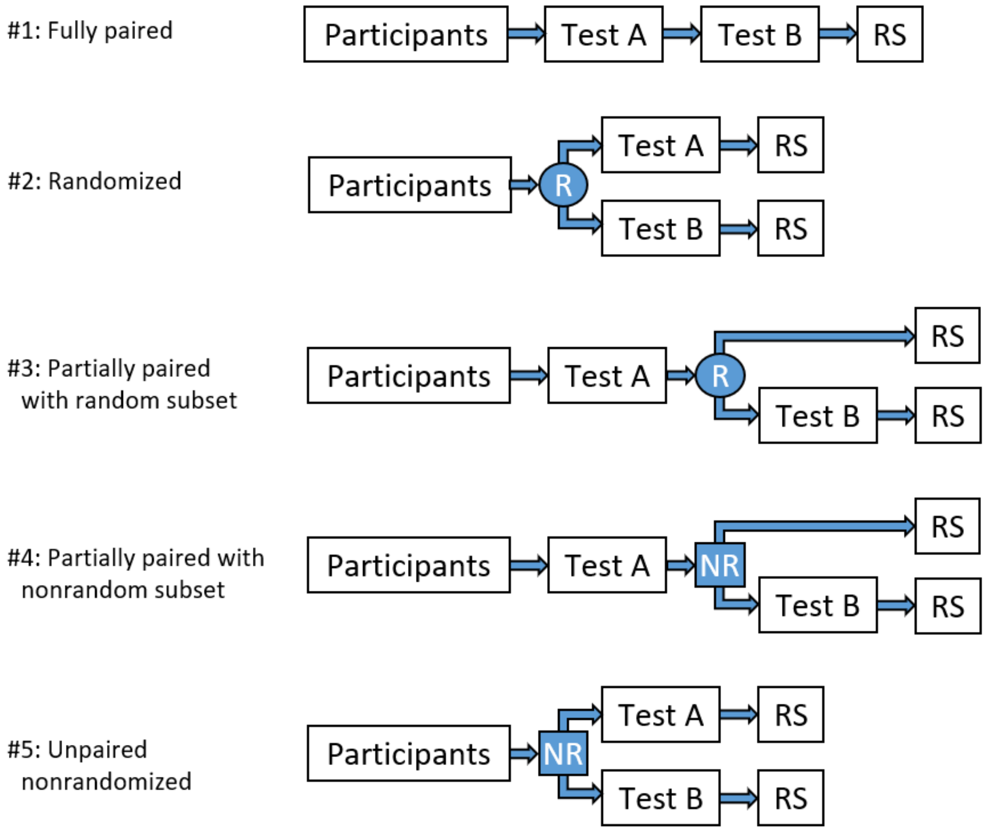
Adapted from (3) under CC BY 4.0. * The examples of risk of bias given for comparative diagnostic test accuracy studies are additional to those given for noncomparative DTA studies.

1.3. Comparative study designs

Knowledge of comparative DTA study designs will help the user to identify relevant studies and assess risk of bias. Here we briefly describe five distinct comparative DTA study designs identified in a review of the literature (2) (Figure 1). The descriptions are based on comparisons of two index tests, but they could equally apply for comparisons of more than two tests.

Comparative DTA studies in which each participant receives both index tests A and B are the most common (#1: fully paired study; the term ‘paired’ is used even when participants receive more than two tests). This design ensures that the participants receiving index test A and participants receiving index test B are identical in terms of factors affecting test accuracy (confounding factors). However, under certain circumstances, such a design may not be feasible or desirable: for example, if it is unethical for a participant to receive both index tests, or if one index test affects the performance of the other index test. These problems can be overcome if each participant can instead be randomized to receive either one index test or the other (#2: randomized study). If the sample size is sufficiently large, randomization is expected to produce index test groups of participants that are comparable in terms of confounding factors. Therefore, fully paired and randomized studies are considered to be the most robust comparative DTA study designs.

Figure 1. Comparative DTA study designs identified in (2).



Adapted from (2) under CC BY 4.0. Abbreviations: R: random allocation, NR: nonrandom allocation, RS: reference standard. These flow diagrams assume a single gate for participant enrolment and two index tests in the comparison. The partially paired designs (#3 and #4) can be illustrated in different ways, but require that a subset of participants received multiple index tests.

If some participants received only one index test while others received both index tests, we refer to the study as ‘partially paired’. The risk of bias then depends on the reasons or mechanisms why some participants received only one of the index tests. If participants are randomly selected either to receive one index test or to undergo both index tests (#3: partially paired, random subset study), the risk of bias is expected not to differ from that of a randomized study. If, however, a nonrandom mechanism is used to decide whether participants would receive one or both index tests, there is potential for confounding (#4: partially paired, nonrandom subset study). An example would be when the clinician only invites a participant to undergo the second index test because of diagnostic uncertainty after the first index test. Confounding can be avoided by analyzing only the paired subset of participants, if these data are reported; but this subset may not be representative of the target population.

In some studies, participants receive only one of the index tests without randomization (#5: unpaired nonrandomized study). For example, the allocation may be driven by participants' or clinicians' preference, or may be done using a quasi-random method (such as alternation or based on date of birth). Due to the serious potential for confounding in study designs #4 and #5, review authors may want to exclude these designs from the comparative analysis of a systematic review, particularly if a sufficient number of fully paired and/or randomized studies are available.

2. Assessing risk of bias with QUADAS-C

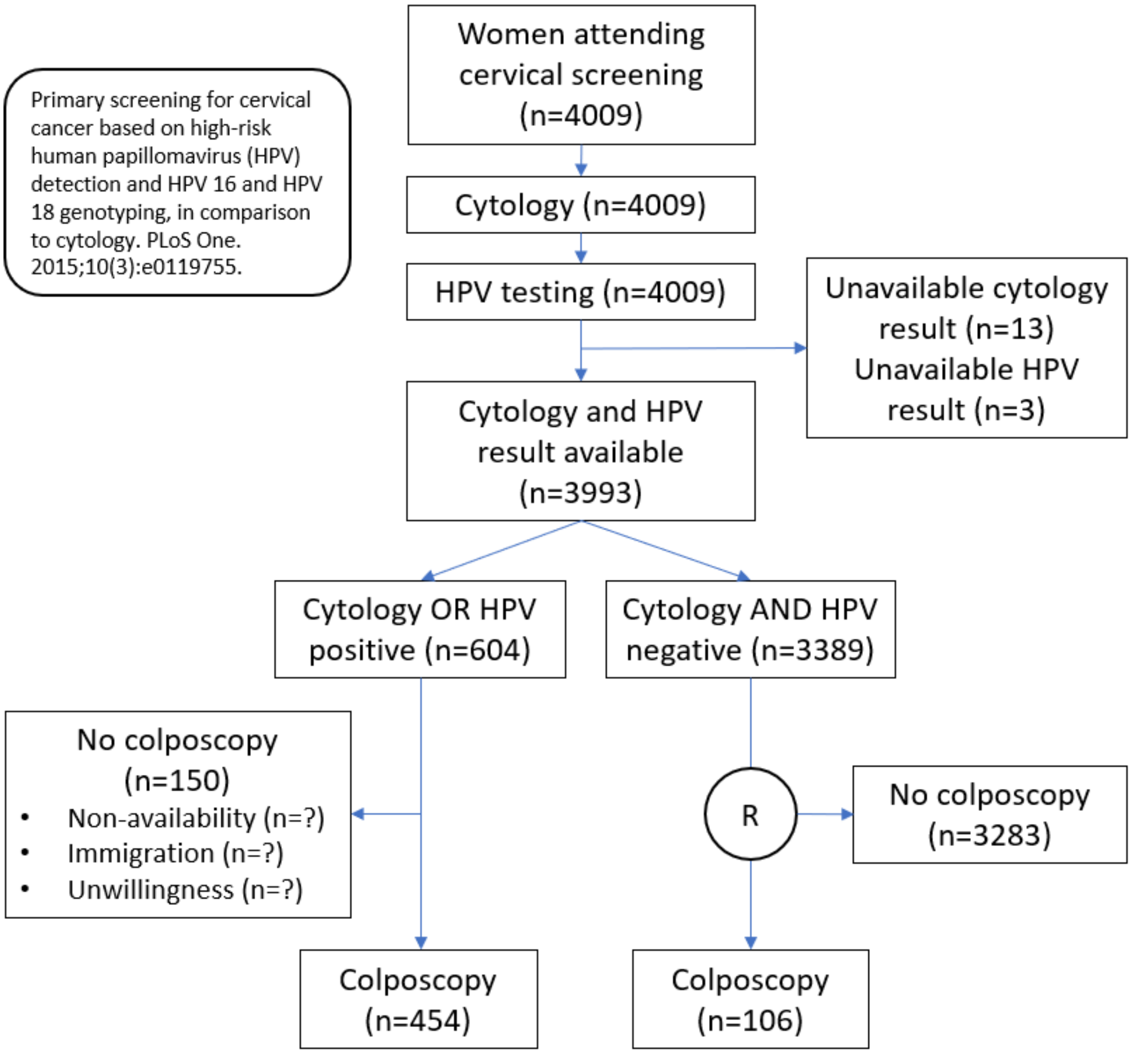
QUADAS-C is completed in four phases, similar to the process for QUADAS-2: 1) stating the review question, 2) tailoring the tool to the review and developing review-specific guidance, 3) reviewing the study flow diagram or constructing one if none is reported, and 4) judging risk of bias. These phases were previously explained in the QUADAS-2 paper (1). Below, we provide an overview of these phases with additional considerations for comparative DTA studies (Table 2).

In the following sections, we expand on the fourth phase, judgment of risk of bias. As stated previously, QUADAS-C only deals with risk of bias, but we strongly recommend that users also assess concerns regarding applicability as part of their assessment. Judgment of risk of bias is composed of three sections: (I) information used to support the risk of bias judgment, (II) signaling questions, and (III) judgment of risk of bias for each domain.

Table 2: The four phases involved in completion of QUADAS-C

1. Stating the review question	Users should clearly specify which index tests will be compared ('what is being compared with what') for which target condition and in which population. If the review consists of several comparative questions, it is important to state them all. An 'index test' may also be a diagnostic strategy consisting of multiple tests. For example, the review aim may be to evaluate whether ultrasonography can be a triage test for computed tomography (CT) to detect acute appendicitis. The comparison is then between CT (index test A) and a diagnostic strategy consisting of ultrasonography and CT (index test B).
2. Tailor the tool and develop review-specific guidance	QUADAS-C should be tailored to each review by adding or omitting signaling questions where needed. Example signaling questions that users might consider adding are provided in Section 2.4. Furthermore, review-specific guidance should be developed on how to answer individual signaling questions and to judge risk of bias in each domain. We recommend piloting the tailored tool and refining the review-specific guidance if inter-rater agreement is insufficient.
3. Flow diagram	We recommend users to review the published participant flow diagram for each study or to construct one if none is reported or the published diagram is inadequate. The flow diagram of a comparative diagnostic test accuracy study should ideally include information on how participants were recruited, how participants were assigned to index tests, the sequence of index tests performed on each participant (if the study is paired), and the number of participants that were missing or did not receive a reference standard. See Figure 2 for an example.
4. Judgment of risk of bias (and concerns regarding applicability)	See following sections (§2.1 to 2.3) for the judgment of risk of bias. Concerns regarding applicability are not part of the QUADAS-C tool but we strongly recommend that users make this judgment for the comparison based on the answers to the QUADAS-2 tool.

Figure 2. Participant flow diagram for a fully paired study comparing cytology and human papillomavirus (HPV) testing for cervical cancer screening. The reference standard is colposcopy with or without biopsy. 'R' indicates that a random subset of women were selected to undergo colposcopy.


Primary screening for cervical cancer based on high-risk human papillomavirus (HPV) detection and HPV 16 and HPV 18 genotyping, in comparison to cytology. PLoS One. 2015;10(3):e0119755.
Women attending cervical screening (n=4009)
Cytology (n=4009)
HPV testing (n=4009)
Unavailable cytology result (n=13)
Unavailable HPV result (n=3)
Cytology and HPV result available (n=3993)
Cytology OR HPV positive (n=604)
Cytology AND HPV negative (n=3389)
No colposcopy (n=150)
• Non-availability (n=?)
• Immigration (n=?)
• Unwillingness (n=?)
R
No colposcopy (n=3283)
Colposcopy (n=454)
Colposcopy (n=106)

2.1. Information used to support the risk of bias judgment

When using QUADAS-2, users are invited to record the information used to reach the risk of bias judgment for each domain. The QUADAS-C extension requires additional information to be recorded relevant to test comparisons (Table 3).

Users should record the type of comparative DTA study design. Identifying the study's design is key in recognizing its general strengths and limitations, and helps determine which signaling questions are not applicable. For some studies, the design may not be apparent until users have constructed a flow diagram or have answered the signaling questions. Users may select one of the five study designs described in Section 1.3 or choose 'other' and describe the specific study design in detail. As noted in Section 1.2, QUADAS-C is designed for assessment of risk of bias in study designs #1 and #2, but users may find it helpful to also complete the tool for other study designs.

Users should also record additional information relevant to the comparison of index tests. Free text fields at the beginning of each domain can be used for this purpose. For example, users can describe whether each study participant underwent all index tests (Patient Selection domain), the randomization process for participants who were randomized to index tests (Patient Selection domain), and the order of index tests undergone by each participant in paired studies (Index Test domain). Furthermore, due to potential for disease progression and/or treatment initiation, it may be important to note whether index tests were conducted at the same time or within a specified appropriate time interval (Flow and Timing domain).

Table 3. Information used to support risk of bias judgments in QUADAS-2 and QUADAS-C. Information required by QUADAS-C is italicized.

Domain	Text prompting supporting information
Comparative study design	*Which of the following study designs does the primary study most strongly resemble?* - *#1 Fully paired* - *#2 Randomized* - *#3 Partially paired with random subset* - *#4 Partially paired with nonrandom subset* - *#5 Unpaired nonrandomized* - *Other (please describe the study design)*
Patient Selection	Describe methods of patient selection. Describe included patients (previous testing, presentation, intended use of index test, and setting). *Describe how patients were allocated to receive each of the index tests. If randomization was used to assign individual patients (or clusters of patients) to index tests, describe the randomization process.*
Index Test	Describe the index tests and how they were conducted and interpreted. *For paired comparative studies, describe the order in which the index tests were performed.*
Reference Standard	Describe the reference standard, how it was conducted and interpreted, *and whether any of the index tests were part of the reference standard.*

Flow and Timing	Describe any patients who did not receive the index tests or reference standard or who were excluded from the analysis. Describe the time interval and any interventions between the index tests and the reference standard. *Describe the time interval and any interventions between the index tests being compared.*

2.2. Guidance on how to answer signaling questions in QUADAS-C

Following recording of the information above (Table 3), users need to answer a number of signaling questions relating to each of the four domains. In the following sections, we provide an explanation for each signaling question, accompanied by examples from primary studies. A 'living' library of reference examples for QUADAS-C items is being developed, and can be found on https://osf.io/hq8mf/files/.

2.2.1. Domain: Patient Selection

Was the risk of bias for each index test judged 'low' for this domain? (C1.1)

If the accuracy estimates of one or more index tests are considered to be at high risk of bias, their comparison will also be at high risk of bias. Therefore, each domain in QUADAS-C starts with the question whether the risk of bias for that domain for each index test (as judged using QUADAS-2) is 'low'. If one or more index tests in the comparison is judged 'unclear' or 'high' in a QUADAS-2 domain, this question should be answered 'no'.

Ideally, test comparisons should be made in participants who are suspected of having the target condition, who are consecutively (or randomly) enrolled, through appropriate in/exclusion criteria. If not, there is a high risk that the estimate of comparative accuracy is biased.

As with all signaling questions, a 'no' answer signals that there may be a problem with this domain, but this should not automatically lead to a 'high risk of bias' judgment. For example, if index test A was judged to be at 'low risk of bias' and index test B at 'unclear risk of bias' for this domain, the current signaling question would be answered 'no' as a result of insufficient information to judge the risk of bias of index test B. Despite answering 'no', some users may decide that this is not enough to judge the QUADAS-C domain as 'high risk of bias'.

There may be cases where the accuracy estimates for each test may be biased, but the comparative accuracy estimate may not be biased. This may be the case if the bias for each test would have the same direction as well as the same magnitude on the relevant scale (absolute or relative difference). However, predicting the direction and magnitude of bias remains difficult and judging 'low risk of bias' in QUADAS-C based on such predictions requires careful motivation. The same applies to identical signaling questions (C2.1, C3.1, and C4.1) in other domains.

Example of a 'No' answer:
In a comparative DTA study comparing magnetic resonance imaging (MRI) and magnetic resonance arthrography (MRA) for the diagnosis of superior labrum anterior posterior (SLAP) lesions of the shoulder using routinely collected data, participants were included only if they had undergone MRI, MRA, and the reference standard, which was surgery. As participants who received only MRI, received only MRA, or did not receive the reference standard are likely to have been excluded, this study group could be a highly selected subsample of all participants who are suspected of the target condition. The QUADAS-2 question '*Did the study avoid inappropriate exclusions?*' was answered 'no' for both MRI and MRA and thus the risk of bias was judged to be 'high' for each index test in QUADAS-2. Since at least one of the index tests is judged to be at 'high risk of bias' in QUADAS-2, the current QUADAS-C question was answered 'no'. (5)

Was a fully paired or randomized design used? (C1.2)

This question asks whether the study used either a design in which each participant received all index tests (#1: fully paired study) or a design in which participants were randomly allocated to one of the index tests (#2: randomized study). Both designs avoid confounding by ensuring like-with-like comparisons. If neither design was used (e.g., a study in which the clinician decided whether participants received index test A or B) there is a high risk of bias. An important exception is a #3: partially paired, random subset design. Like #1 and #2, this design should protect against confounding and may imply a ‘low risk of bias’ judgment for this domain despite a ‘no’ answer.

With the exception of design #3, answering ‘no’ to this question should almost always prompt a ‘high risk of bias’ judgment for this domain. Users should be cautious when interpreting studies that have a 'no' to this question and should not trivialize a 'high risk of bias’ judgment for this domain, even if all other domains are judged 'low’. Users could even consider using this question as a screening tool for not completing the rest of QUADAS-C, as a criterion for excluding studies from the primary comparative meta-analysis, or as a criterion for conducting a sensitivity analysis.

Special consideration is needed for retrospective studies that present fully paired data. Such studies ensure like-with-like comparisons, but participants may be selected in a way that they are unlikely to be representative of the target population. This is the case if investigators only included participants who were eligible for undergoing all index tests (e.g., those with greater diagnostic uncertainty) and excluded participants who only received one of the tests. When assessing such studies, we recommend answering ‘yes’ to this signaling question but to judge the risk of bias appropriately using signaling questions in QUADAS-2 (see example in C1.1).

It is possible that study investigators intended to conduct a fully paired study, but in some participants the results of one index test were missing. In this case, this question should be answered ‘yes’ but the question regarding missing data (C4.4 in Flow and Timing domain) may be answered ‘no’.

Example of a ‘No’ answer:
In a comparative DTA study comparing MRI and MRA for the diagnosis of shoulder SLAP lesions, participants received either MRI or MRA based on their physicians' request. (6)

Was the allocation sequence random? (C1.3) *– only applicable to randomized designs*

This question was adapted from the revised Cochrane risk of bias tool for randomized trials (7). If randomization is used to assign participants to one index test or another, the randomization process needs to be scrutinized. Random allocation sequences include computer generated random numbers, random number tables, and drawing lots. Non-random allocation sequences include alternation, methods based on dates (e.g., birth or admission), and decisions made by clinicians or patients (7).

Example of a ‘No’ answer:
A study compared two percutaneous liver biopsy needles (Menghini and Tru-cut) for the diagnosis of liver cirrhosis (verified by ‘internationally agreed criteria’). While this study was described as a

randomized study, the investigators used an alternating allocation scheme: *"Randomization was achieved by changing the type of needle at the beginning of each month".* (8)

Was the allocation sequence concealed until patients were enrolled and assigned to index tests? (C1.4) *– only applicable to randomized designs*

This question was also adapted from the revised Cochrane risk of bias tool for randomized trials (7). Appropriate methods to conceal allocation include central randomization schemes (e.g., performed by an independent central pharmacy, or by a telephone or internet-based randomization service provider) and sealed envelopes (opaque, sequentially numbered and opened only after being irreversibly assigned to participants) (7). In the case of cluster randomized comparative DTA studies, additional considerations as described by Eldridge et al. (9) may apply.

Example of a 'No' answer:
In the study comparing the Menghini and Tru-cut needles, the allocation sequence was not concealed as a predictable allocation method (alternating allocation) was used (8).

Example of a 'Yes' answer:
A randomized study comparing conventional white-light imaging endoscopy with narrow-band imaging endoscopy for the diagnosis of gastric mucosal cancers reported:
"Randomization was performed promptly on-site using tables of random numbers stratified by hospital, and the results thereof were kept in sealed, numbered envelopes. The random allocation sequence was prepared at the data management center. Both the assignment result and the corresponding envelope number were recorded by the data management center. At each participating hospital, sealed envelopes were stored by a third party who was not involved in the study, and the envelopes were opened by an assistant physician in serial order only when randomization was performed. The assigned patient identification number, envelope number, and assignment result were recorded on-site and faxed to the data management center on the day of the examination." (10)

2.2.2. Domain: Index Test

Was the risk of bias for each index test judged 'low' for this domain? (C2.1)

As stated previously, if the accuracy estimates of one or more index tests are judged to be at high or unclear risk of bias, their comparison will also be at high or unclear risk of bias. For example, if one of the index tests has been interpreted with knowledge of the reference standard results, the accuracy of that index test may appear to be higher than that of other index tests, even in the absence of a true difference in accuracy. Even if all index tests in the comparison are interpreted with knowledge of the reference standard results, it is unlikely that the bias will affect each index test with the same magnitude. Therefore, the comparison could be biased.

Example of a 'No' answer:
A study compared the QuickC6 antibody test with the combined antibody test EnzygnostG + DakoM for the diagnosis of acute Lyme neuroborreliosis. The threshold for the QuickC6 test was selected using the study data, a method that is known to overestimate performance compared

with the use of a pre-defined threshold. The thresholds for EnzygnostG + DakoM tests were based on the manufacturer's recommendation. (11)

The QUADAS-2 question 'If a threshold was used, was it pre-specified?' was answered 'no' for QuickC6 and thus the risk of bias was judged to be 'high'. For EnzygnostG + DakoM this was answered 'yes' and risk of bias was judged to be 'low'. Since at least one of the index tests is judged to be at 'high risk of bias' in QUADAS-2, the current QUADAS-C question was answered 'no'.

Were the index test results interpreted without knowledge of the results of the other index test(s)? (C2.2) *– only applicable if patients received multiple index tests*

This question is only applicable to fully paired and partially paired studies (e.g., #1, #3, and #4). Index tests should be interpreted without knowledge of (or blind to) the results of other index tests. Three considerations should be made when judging the risk of bias:

1. The risk of bias will depend on the degree of subjectivity involved in interpreting the test result. Estimates for an index test that requires subjective interpretation (e.g., whether or not a tumor can be seen on a radiograph) will be more susceptible to bias than estimates for an index test that produces an unambiguous output (e.g., a quantitative glucose measurement with a prespecified positivity threshold).
2. The risk of bias will depend on the order in which the index tests are carried out and interpreted. Suppose we are comparing two index tests, A and B. If A is always carried out and interpreted prior to B being performed, then we can infer that A is always interpreted without knowledge of the results of B. Then we would only need B to be blinded to the results of A.
3. A 'no' answer will not always imply high risk of bias if an index test is compared with a test strategy consisting of multiple index tests. Consider a study where we compare ultrasound (US) against a combination of US and CT for appendicitis. The person interpreting CT results does not have to be blinded to US results if, in clinical practice, US is usually done before CT and US results are usually available to the CT interpreter. Only the US needs to be interpreted without knowledge of CT results.

Example of a 'No' answer:
A study compared the accuracy of the Mini-Mental State Examination (MMSE) and the General Practitioner Assessment of Cognition (GPCOG) tests for dementia screening in the same participants. The tests were administered by the same nurse in the same session (MMSE followed by GPCOG), and thus the nurse was aware of the results of the MMSE when administering the GPCOG. (12)

Is undergoing one index test <u>unlikely</u> to affect the performance of the other index test(s)? (C2.3) *– only applicable if patients received multiple index tests*

This question is only applicable to fully paired and partially paired studies (e.g., #1, #3, and #4). In these studies, bias may occur if undergoing one index test will affect or interfere with the performance of a subsequent index test. This question is different from the previous one (C2.2) which focuses on the bias introduced by the people interpreting the test results, whereas the current question covers other unwanted effects that an index test may have on subsequent index tests. Examples include: participants experiencing learning effects or fatigue when completing multiple questionnaires; a second biopsy needle being used in tissue already distorted by the first needle; and a second blood marker test not having enough blood sample left after the first. This issue is similar to the issue of carryover effects in crossover trials of interventions (13). Refusal or inability to receive the second index test after undergoing the first will lead to missing data; this issue can be addressed in the Flow and Timing domain (question C4.4) instead.

Some studies may vary the order of index tests (e.g., by randomizing the order) with the intention of mitigating against bias introduced by such effects. However, randomizing the order does not necessarily prevent learning effects or other order-related effects. For instance, learning effects on participants may result in two index tests appearing more similar in performance, regardless of whether the order was randomized or not. If test A affects the performance of test B, randomizing the order can theoretically lead to a low risk of bias in the comparison if 1) test B also affects test A if the order is reversed, and 2) the resulting bias for each test has the same direction and the same magnitude on the relevant scale (absolute or relative difference). Ideally, if one index test is likely to influence the performance of subsequent tests, each participant should receive only one index test.

Example of a ‘No’ answer
A study compared the accuracy of three cognitive tests (MMSE, Rowland Universal Dementia Assessment Scale, and the modified Kimberley Indigenous Cognitive Assessment) for detection of dementia in the same participants. Since these cognitive tests contain similar components (e.g., drawing and hand exercise tasks), participants may show some learning effects. (14)

Were the index tests conducted and interpreted without advantaging one of the tests? (C2.4)

In a fair comparison, all index tests should be performed and interpreted under similar circumstances to avoid inappropriately advantaging one of the tests. If one of the index tests was conducted or interpreted in a way that is remarkably different from the other index test(s) – for example, one biomarker assay was conducted using fresh samples, whereas a competing biomarker assay was conducted using frozen samples – then there is potential for bias. However, differences may be acceptable if they reflect clinical practice. This question is intended to cover differences in test conduct and interpretation other than those already captured by the previous two questions (C2.2 and C2.3).

Example of a ‘No’ answer:
In an unpaired study comparing CT to reduced-dose CT for the diagnosis of pediatric appendicitis, the ‘normal’ CT group was assessed using old generation CT scanners, while the reduced-dose CT group was assessed using modern CT scanners. (15)

2.2.3. Domain: Reference Standard

Was the risk of bias for each index test judged 'low' for this domain? (C3.1)

If the reference standard is likely to misclassify participants, not only the accuracy of individual tests is at risk of bias, but also their comparison. Furthermore, the reference standard should preferably be interpreted without knowledge of any of the index tests.

> Example of a 'No' answer:
> In a study comparing MRI and MRA for the diagnosis of superior labrum anterior posterior (SLAP) lesions, the operating surgeons (the reference standard was open or arthroscopic surgery) were aware of the findings of both index tests. The QUADAS-2 question '*Were the reference standard results interpreted without knowledge of the results of the index test?*' was answered 'no' for both index tests. As one type of imaging may be preferred in planning the surgical approach, the risk of bias was judged to be 'high' for both index tests in QUADAS-2. As neither test was at low risk of bias, the current QUADAS-C question was answered 'no'. (16)

Did the reference standard avoid incorporating any of the index tests? (C3.2)

'Incorporation' means that an index test is part of the reference standard. If this is the case, we can expect the agreement between the index test and the reference standard to increase, leading to a superficially higher accuracy for the index test. There is a clear risk of bias if one index test is part of the reference standard, but another index test is not. Even when all index tests are part of the reference standard, there could be differences in the weight or contribution of each index test to the final diagnosis. Ideally, none of the index tests should be part of the reference standard. This question is not about whether the reference standard results were interpreted without knowledge of the index test results, which is addressed in the Reference Standard domain of QUADAS-2.

> Example of a 'No' answer:
> A study compared the accuracy of two questionnaires (MMSE and GPCOG) for detection of dementia, which was detected by the reference standard Cambridge Cognitive Examination (CAMCOG) criteria. One of the index tests (MMSE) was a subtest within the larger CAMCOG assessment. (12)

2.2.4. Domain: Flow and Timing

Was the risk of bias for each index test judged 'low' for this domain? (C4.1)

As with the previous three domains, bias in the accuracy of the individual tests may also bias the comparison. The timing between each index test and the reference standard should be appropriate (e.g., the target condition should not change in the meanwhile), all participants should receive the same reference standard, and all participants should be included in the analysis.

If some study participants do not receive the reference standard, the *relative* accuracy of tests can sometimes be correctly estimated even if the accuracy of each individual test cannot be estimated. For a discussion on this topic, see Box.

Example of a ‘No’ answer:
In a fully paired study comparing cytology with HPV testing for cervical cancer screening (Figure 2), 3433/4009 (85.6%) women did not receive the reference standard and were excluded from the analysis (150 women who were positive for at least one index test and 3283 randomly selected women who were negative for both tests). The QUADAS-2 questions ‘*Did all patients receive a reference standard?*’ and ‘*Were all patients included in the analysis?*’ were answered ‘no’ for both index tests and the risk of bias was judged ‘high’ for both index tests in QUADAS-2. Therefore, the current QUADAS-C question was answered ‘no’. (17)

Box: Valid relative accuracy estimates despite partial verification

Imagine a comparative DTA study in which participants with at least 1 positive index test result undergo the reference standard, while participants with only negative results do not. In this scenario, even though the accuracy of each index test cannot be estimated, some comparative accuracy measures are unbiased. An example is *relative sensitivity* (sensitivity A/sensitivity B) and *relative false positive rate* (false positive rate A/false positive rate B). These measures require only the true positives and false positives of each index test to be known; for example, relative sensitivity can be estimated through the comparison of detection rates (18). This method assumes that the prevalence of the target condition is similar across index test groups (as in fully paired or randomized studies) and any missing data is missing completely at random.

Was there an appropriate interval between the index tests? (C4.2)

Generally, the time interval between index tests can be considered appropriate when all index tests were performed at the same time after enrolment, to exclude the possibility of disease progression or change in patient management. However, what is considered ‘appropriate’ may differ significantly between target conditions and the tests being evaluated. For example, a time interval of a few days may be acceptable for a slowly progressive disease in contrast to an acute and rapidly progressive disease. Performing index tests simultaneously is unnecessary (perhaps even undesirable) when comparing index tests that are usually conducted at different timepoints, reflecting clinical practice.

Example of a ‘No’ answer:
A fully paired study compared the accuracy of chest ultrasound with that of chest CT to detect parenchymal abscess or necrosis in children with pneumonia complicated by parapneumonic effusion. The mean time interval between the two tests was 2.7 days (range 0 to 8 days); the order of tests was unreported. The delay in testing may have led to development and better visualization of lesions on imaging. (19)

Was the same reference standard used for all index tests? (C4.3)

While the QUADAS-2 question ‘*Did all participants receive the same reference standard?*’ asks whether the same reference standard was used for all participants *within* an index test group (i.e. group of participants that receive the same index test), this QUADAS-C question asks whether the same reference standard was used *across* index test groups. If different reference standards are applied to those receiving index test A (e.g., surgery) and to those receiving index test B (e.g., followup), then the comparison may be biased. This problem is easiest to imagine for unpaired or partially paired studies, where different participants can receive different reference standards. However, #1: fully paired studies can also suffer from this problem, as illustrated by the example below.

If studies use a *single* reference standard to verify results from all index tests, this question should be answered 'yes'. If different reference standards were used across index test groups, users should answer 'no' but consider whether the reference standards are exchangeable (i.e. detect the same target condition, in the same way). If the reference standards can be considered to be exchangeable, a 'no' answer to this question should not imply a high risk of bias.

Users may encounter unpaired or partially paired studies that use the same two reference standards across all index test groups, in which participants with a positive index test result receive reference standard A, and participants with a negative index test result receive reference standard B. This signaling question will be answered 'yes' for such studies, but the proportion of participants who receive reference standard A and reference standard B will most likely differ across index test groups. This problem is not addressed by the current signaling question. However, the aforementioned QUADAS-2 question *'Did all patients receive the same reference standard?'* will flag such studies as potentially problematic.

Example of a 'No' answer:
In a fully paired study comparing three fecal occult-blood tests (Hemoccult II, Hemoccult II Sensa, and HemeSelect) for colorectal cancer screening, participants testing positive for Hemoccult II and/or HemeSelect were recommend to receive the reference standard colonoscopy. However, those testing solely positive for Hemoccult II Sensa (but negative on the other two index tests) were suggested to undergo sigmoidoscopy and repeated Hemoccult II testing at 6 and 12 months, rather than the preferred reference standard colonoscopy. (20)

Are the proportions and reasons for missing data similar across index tests? (C4.4)

This question is adapted from the Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) tool (21). Missing data occurs if test results are unavailable, invalid, valid but inconclusive (22), or if participants are excluded from the analysis. Users should carefully consider whether the proportion of and reasons for missing data are likely to have an impact on the test comparison. Unlike studies of interventions, missing data are rarely imputed in diagnostic accuracy studies. The best method of avoiding bias due to missing data is to avoid missing data altogether during the design and execution of a study (23).

Example of a 'No' answer:
In a fully paired study comparing visual inspection with acetic acid (VIA) with HPV testing for cervical cancer screening, 30/4039 (0.7%) women were excluded from the VIA group, and 493/4039 (12.2%) women were excluded from the HPV group. The authors reported that the women were excluded because of 'incomplete or inconclusive reference investigations', but did not explain the difference in the proportion of missing participants. (24)

Examples of a 'Yes' answer:

In a randomized study comparing conventional white-light imaging (C-WLI) endoscopy (n=180) versus the magnifying endoscopy with narrow-band imaging (M-NBI) (n=182) for diagnosis of gastric mucosal cancer, the authors described the missing data as follows:

"Four patients in the C-WLI group (one patient's lesion was 10 mm in diameter, one was discontinued from the examination because of Mallory–Weiss syndrome, and 2 had a missed biopsy) and 5 patients in the M-NBI group (one was examined with an unpermitted endoscope and 4 missed biopsy) were excluded. Data for 176 patients in the C-WLI group and 177 patients in the M-NBI group were used for the final analysis."

We answered 'yes' because of the small proportion of participant loss and reasons provided for missing data (in particular 'missed biopsies') being similar between the two groups. (10)

2.3. Judgment of risk of bias for each domain

Similar to QUADAS-2, after answering the signaling questions, users are invited to judge risk of bias in QUADAS-C as 'low', 'high' or 'unclear' for each domain. In QUADAS-C, the judgment of risk of bias in the comparison can be entirely based on the signaling questions within QUADAS-C. The starting question of each domain *('Was the risk of bias for each index test judged 'low' for this domain?'*) is designed to summarize the risk of bias information captured by QUADAS-2.

Users should carefully develop review-specific guidance on how to move from signaling question answers to a judgment of risk of bias for each domain.

- Generally, if the answer is 'yes' to all signaling question for a domain, the risk of bias can be judged 'low'.
- A 'no' answer should lead to a 'high risk' judgment, if the bias introduced by the design feature is of such concern that the entire domain is deemed problematic. For example, a 'no' answer to the signaling question *'Was a fully paired or randomized design used?'* should almost always prompt a 'high risk of bias' judgment for the Patient Selection domain (except when a #3: partially paired, random subset design is used).
- The 'unclear' option does not imply 'moderate' risk of bias, but rather that there is insufficient information to judge risk of bias.

2.4. Adding signaling questions to tailor the tool

As previously noted (Table 2), a crucial step in completing QUADAS-C is to tailor the tool to each review by modifying, adding, or omitting signaling questions. Adding new signaling questions should only be considered if only there is strong reason to do so, to avoid complicating the tool. Any new signaling questions should focus on aspects of methodological quality rather than quality of reporting. When adding new signaling questions, we recommend phrasing them in such a way that a 'yes' answer implies 'low risk of bias' in order to avoid potential confusion when judging risk of bias.

During QUADAS-C's development process, we identified some issues suggestive of potential bias but not considered common enough to be addressed in the core QUADAS-C tool. Signaling questions addressing these issues are listed in Table 4. Users may use these signaling questions to tailor the tool, if the issues addressed by these questions are particularly important or frequent in their systematic review.

Some users may be interested in addressing conflict of interest issues in QUADAS-C. While it is possible to add signaling questions regarding conflict of interest, these concerns can also be recorded elsewhere. TACIT (Tool for Addressing Conflicts of Interest in Trials) is a tool currently in development that provides review authors with a framework for addressing conflict of interest issues in studies; its release is planned for 2022 (http://tacit.one).

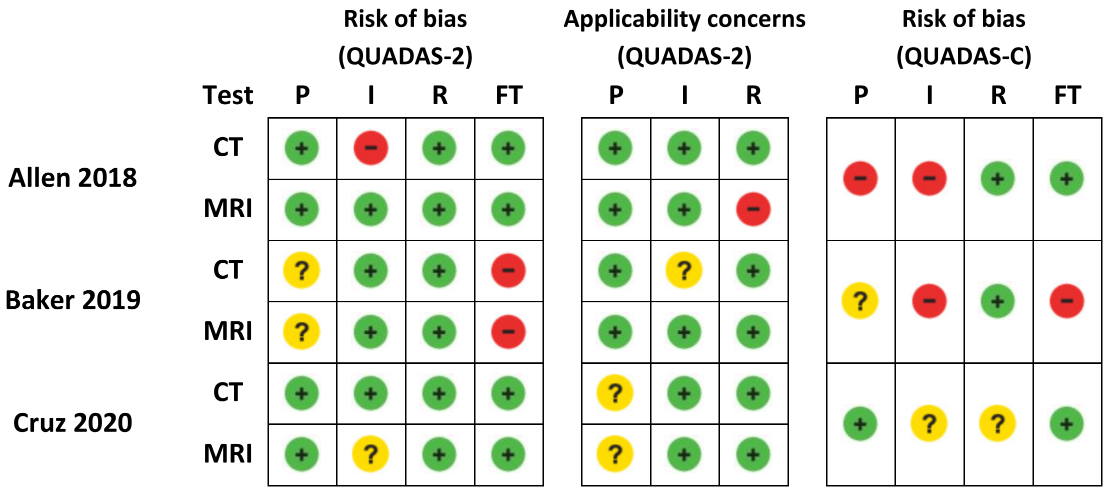
Table 4. Signaling questions that were not selected for the core QUADAS-C tool but users can use to tailor their tool.

Domain	**Signaling question (risk of bias)**	**Applicable to**
Patient Selection	Were the same patient selection criteria used for those assigned to each index test?	Unpaired nonrandomized studies
Patient Selection	If patients received all index tests, was the decision to use all index tests made before participants were recruited?	Paired studies
Patient Selection	Did the study avoid using prior tests as inclusion criteria that were correlated with only one of the index tests?	All
Index Test	Did the study avoid using index test thresholds that are likely to advantage some of the index tests?	Studies that assess nonbinary tests
Reference Standard	Is the mechanistic basis of one index test more closely shared with the reference standard than the other index test(s)?	All

3. Presenting QUADAS-C results in systematic reviews

We recommend systematic reviews present QUADAS-2 as well as QUADAS-C results for comparative DTA studies. The QUADAS-C results are specific to a particular test comparison in a study. If a review includes more than one test comparison, QUADAS-C results should be presented for each comparison. Below we present tabular and graphical suggestions on how to present QUADAS-2 and QUADAS-C results together. Tables A, B, and C display the results per study; figures A and B display proportion of studies with low, high, or unclear judgments. Templates are available at www.quadas.org.

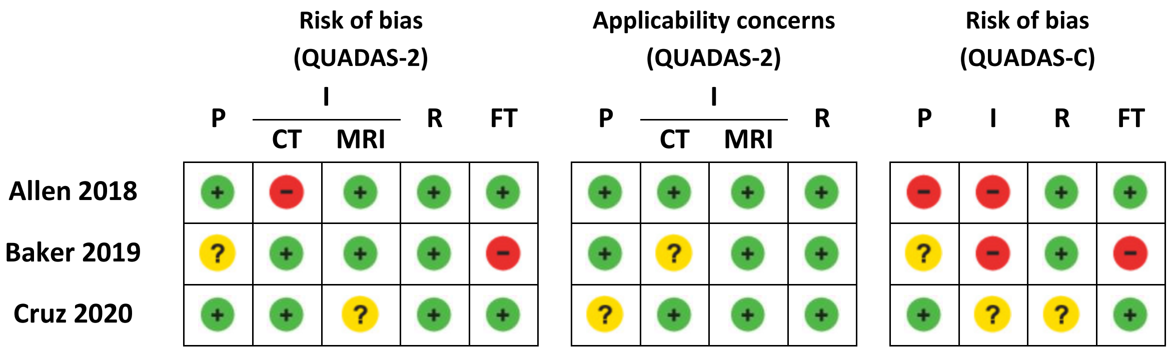
Table A: Example with QUADAS-2 judgments to the left and QUADAS-C judgments to the right.



		Risk of bias (QUADAS-2)				Applicability concerns (QUADAS-2)			Risk of bias (QUADAS-C)			
	Test	P	I	R	FT	P	I	R	P	I	R	FT
Allen 2018	CT	+	−	+	+	+	+	+	−	−	+	+
	MRI	+	+	+	+	+	+	−				
Baker 2019	CT	?	+	+	−	+	?	+	?	−	+	−
	MRI	?	+	+	−	+	+	+				
Cruz 2020	CT	+	+	+	+	?	+	+	+	?	?	+
	MRI	+	?	+	+	?	+	+				

Abbreviations: P: Patient Selection, I: Index Test, R: Reference Standard, FT: Flow & Timing, CT: computed tomography, MRI: magnetic resonance imaging. + indicates low risk, - indicates high risk, ? indicates unclear risk.

Table B: Table A simplified. This format may be useful if the QUADAS-2 judgments for Patient Selection, Reference Standard, and Flow & Timing domains are the same for each index test. This could also be applied to Table C, Figure A and Figure B.



	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
		CT	MRI				CT	MRI					
Allen 2018	+	−	+	+	+	+	+	+	+	−	−	+	+
Baker 2019	?	+	+	+	−	+	?	+	+	?	−	+	−
Cruz 2020	+	+	?	+	+	?	+	+	+	+	?	?	+

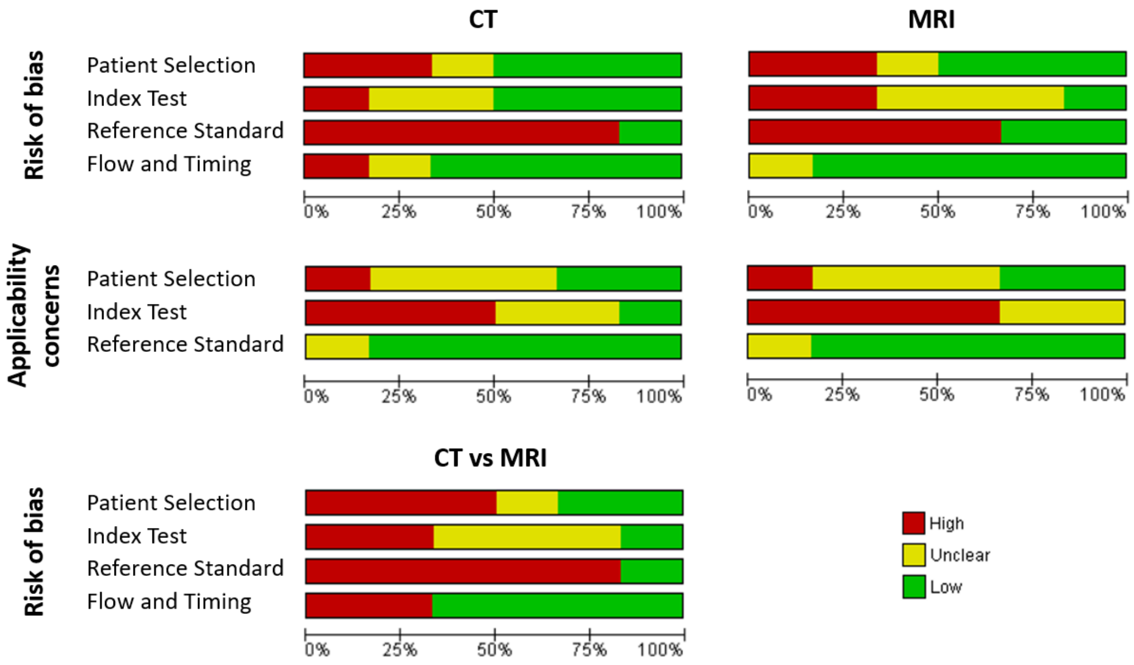
Abbreviations: P: Patient Selection, I: Index Test, R: Reference Standard, FT: Flow & Timing, CT: computed tomography, MRI: magnetic resonance imaging. + indicates low risk, - indicates high risk, ? indicates unclear risk.

Table C: Example with QUADAS-2 judgments on top and QUADAS-C judgments on the bottom. This format may be useful if the review did not include many comparative diagnostic test accuracy studies.

		Risk of bias				Applicability concerns		
	Test	P	I	R	FT	P	I	R
Allen 2018	CT	+	-	+	+	+	+	+
Baker 2019	CT	?	+	+	-	+	?	+
Cruz 2020	CT	+	+	+	+	?	+	+
Allen 2018	MRI	+	+	+	+	+	+	-
Baker 2019	MRI	?	+	+	-	+	+	+
Cruz 2020	MRI	+	?	+	+	?	+	+
Allen 2018	CT vs MRI	-	-	+	+			
Baker 2019	CT vs MRI	?	-	+	-			
Cruz 2020	CT vs MRI	+	?	?	+			

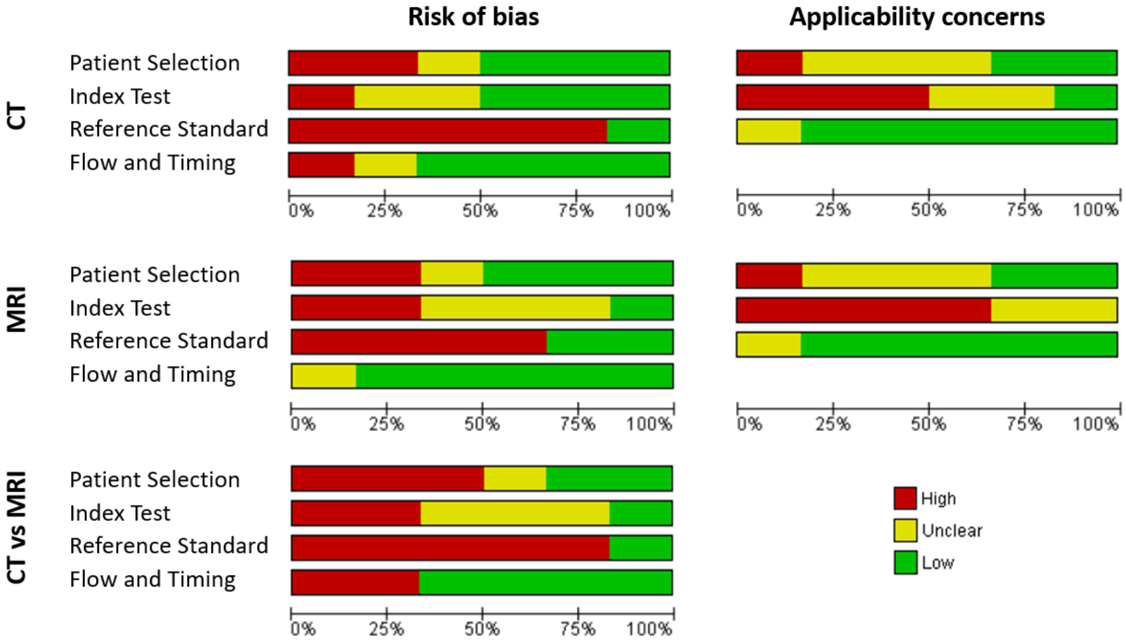
Abbreviations: P: Patient Selection, I: Index Test, R: Reference Standard, FT: Flow & Timing, CT: computed tomography, MRI: magnetic resonance imaging. + indicates low risk, - indicates high risk, ? indicates unclear risk.

Figure A: Example with index tests as columns and the test comparison at the bottom.



Abbreviations: CT: computed tomography, MRI: magnetic resonance imaging

Figure B: Example with index tests as rows and the test comparison at the bottom. This format may be useful if there are many index tests and multiple comparisons in the review.



Abbreviations: CT: computed tomography, MRI: magnetic resonance imaging

4. References

1. 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.

2. 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.

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. 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.

5. Magee T. 3-T MRI of the shoulder: Is MR arthrography necessary? Am J Roentgenol. 2009;192(1):86–92.

6. Tuite MJ, Rutkowski A, Enright T, Kaplan L, Fine JP, Orwin J. Width of high signal and extension posterior to biceps tendon as signs of superior labrum anterior to posterior tears on MRI and MR arthrography. Am J Roentgenol. 2005;185(6):1422–8.

7. 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.

8. Colombo M, Del Ninno E, de Franchis R, De Fazio C, Festorazzi S, Ronchi G, et al. Ultrasound-assisted percutaneous liver biopsy: Superiority of the Tru-Cut over the Menghini needle for diagnosis of cirrhosis. Gastroenterology. 1988;95(2):487–9.

9. Eldridge SM, Campbell MK, Campbell MJ, Drahota AK, Giraudeau B, Reeves BC, et al. Revised Cochrane risk of bias tool for randomized trials (RoB 2) Additional considerations for cluster-randomized trials (RoB 2 CRT) [Internet]. 2020. Available from: https://www.riskofbias.info/welcome/rob-2-0-tool/rob-2-for-cluster-randomized-trials

10. Ezoe Y, Muto M, Uedo N, Doyama H, Yao K, Oda I, et al. Magnifying Narrowband Imaging Is More Accurate Than Conventional White-Light Imaging in Diagnosis of Gastric Mucosal Cancer. Gastroenterology. 2011 Dec;141(6):2017-2025.e3.

11. Skarpaas T, Ljøstad U, Søbye M, Mygland Å. Sensitivity and specificity of a commercial C6 peptide enzyme immuno assay in diagnosis of acute Lyme neuroborreliosis. Eur J Clin Microbiol Infect Dis. 2007 Aug 21;26(9):675–7.

12. Brodaty H, Connors MH, Loy C, Teixeira-Pinto A, Stocks N, Gunn J, et al. Screening for Dementia in Primary Care: A Comparison of the GPCOG and the MMSE. Dement Geriatr Cogn Disord. 2016;42(5–6):323–30.

13. Sibbald B, Roberts C. Understanding controlled trials: Crossover trials. BMJ. 1998 Jun 6;316(7146):1719–20.

14. Radford K, Mack HA, Draper B, Chalkley S, Delbaere K, Daylight G, et al. Comparison of Three Cognitive Screening Tools in Older Urban and Regional Aboriginal Australians. Dement Geriatr Cogn Disord. 2015;40(1–2):22–32.

15. Didier RA, Vajtai PL, Hopkins KL. Iterative reconstruction technique with reduced volume CT dose index: diagnostic accuracy in pediatric acute appendicitis. Pediatr Radiol. 2015 Feb 5;45(2):181–7.

16. Fallahi F, Green N, Gadde S, Jeavons L, Armstrong P, Jonker L. Indirect magnetic resonance arthrography of the shoulder; A reliable diagnostic tool for investigation of suspected labral pathology. Skeletal Radiol. 2013;42(9):1225–33.

17. Agorastos T, Chatzistamatiou K, Katsamagkas T, Koliopoulos G, Daponte A, Constantinidis T, et al. Primary screening for cervical cancer based on high-risk human papillomavirus (HPV) detection and HPV 16 and HPV 18 genotyping, in comparison to cytology. PLoS One. 2015;10(3):e0119755.

18. 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.

19. Kurian J, Levin TL, Han BK, Taragin BH, Weinstein S. Comparison of ultrasound and CT in the evaluation of pneumonia complicated by parapneumonic effusion in children. Am J Roentgenol. 2009;193(6):1648–54.

20. Allison JE, Ekawa IS, Ransom LJ, Adrain AL. A comparison of fecal occult-blood tests for colorectal-cancer screening. N Engl J Med. 1996;334(3):155–9.

21. 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.

22. Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. BMJ. 2013;346:f2778–f2778.

23. Mack C, Su Z, Westreich D. Approaches to Prevent Missing Data. In: Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. Third Edit. Rockville (MD): Agency for Healthcare Research and Quality (US); 2018. Available from: https://www.ncbi.nlm.nih.gov/books/NBK493616/

24. Shastri SS, Dinshaw K, Amin G, Goswami S, Patil S, Chinoy R, et al. Concurrent evaluation of visual, cytological and HPV testing as screening methods for the early detection of cervical neoplasia in Mumbai, India. Bull World Health Organ. 2005;83(3):186–94.