Chapter 8

Assessing risk of bias and applicability

Johannes B. Reitsma, Anne W. Rutjes, Penny Whiting, Bada Yang, Mariska M. Leeflang, Patrick M. Bossuyt and Jonathan J. Deeks

KEY POINTS

- Shortcomings in the design and conduct of test accuracy studies can lead to biased estimates of test accuracy. This is supported by empirical evidence for studies of the accuracy of single tests.
- Cochrane recommends using the QUADAS-2 tool to evaluate the risk of bias and applicability of test accuracy studies.
- QUADAS-2 assesses the risk of bias of a study across four domains: participant selection, index test, reference standard, and flow and timing, as well as an overall assessment of risk of bias. Applicability of the findings to the review question is evaluated for participant selection, index test and reference standard.
- In studies evaluating the accuracy of two or more index tests, an assessment of risk of bias and applicability should be performed for each index test included in the review.
- If the accuracy of different tests will be compared in the review, corresponding comparisons of accuracy in primary studies should also be assessed for risk of bias using QUADAS-C, an extension of QUADAS-2 for comparative accuracy studies.
- Review authors need to provide guidance in both the protocol and the final review on how to answer each signalling question and how to arrive at the judgement of risk of bias for each QUADAS-2 domain.
- Assessments should preferentially be undertaken in parallel by two or more review authors, and there should be an explicit procedure to resolve disagreements.

Cite this chapter as: Reitsma JB, Rutjes A, Whiting P, Yang B, Leeflang MM, Bossuyt PM, Deeks JJ. Chapter 8: Assessing risk of bias and applicability. In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0 (updated July 2023). Cochrane, 2023. Available from https://training.cochrane.org/handbook-diagnostic-test-accuracy/current

8.1 Introduction

The findings of the assessment of risk of bias and applicability play an important role in the systematic review process, in particular in the analysis and interpretation of results. Whether a systematic review of test accuracy studies allows conclusions about the accuracy of a test, or differences in accuracy between two or more tests, will depend on having available studies at low risk of bias that directly apply to the review question.

Certain flaws in the design and conduct of test accuracy studies can produce incorrect or biased results. These flaws put the corresponding study at 'risk of bias', meaning that its internal validity is threatened. Examples of potential issues include the use of an imperfect reference standard, omitting results from the analyses, or interpreting index test results with knowledge of the result of the reference standard.

Even in the absence of such flaws, a test accuracy study may generate results that do not help answer the review question. Study participants may not be similar to those defined by the review question, the test may be used at a different point in the clinical pathway, or the test may be used in a different way than intended in the review question. Such a mismatch may lead to concerns regarding the applicability of the results from a particular study to the review question. In systematic reviews of test accuracy, both the risk of bias and applicability are assessed.

Risk of bias and concerns regarding applicability can be considered at different points in the systematic review process. First, eligibility criteria are defined to ensure that the included studies meet some minimum criteria. For example, a review may include only studies that use a particular reference standard, or include only studies in which a third-generation computed tomography (CT) scanner has been compared to more recent CT scanners. The use of eligibility criteria is discussed in Chapter 5. Detailed assessment of risk of bias and applicability of all studies included in the review is then undertaken and the results reported. This process is addressed in this chapter.

In the analysis phase, studies may be grouped according to characteristics related to risk of bias or concerns regarding applicability. This can be done both in investigations of heterogeneity, which investigate differences in results attributable to identifiable study features, and in sensitivity analyses, which limit the impact of studies of questionable rigour on study estimates, as described in Chapter 9.

The strength of evidence supporting a review's conclusions depends on the overall assessment of risk of bias and applicability of the evidence base. Recommendations for future research are made, noting particular methodological deficiencies in the existing studies, as outlined in Chapter 12. If the results of individual studies are biased and they are synthesized without consideration of this bias, then the conclusions of the review cannot be trusted.

The focus of this chapter is on assessing risk of bias and applicability for individual studies in a review. These studies may have assessed the accuracy of one index test, of multiple tests, or they may have compared the accuracy of two or more tests.

Specific issues arise for individual studies making a comparison between two or more index tests. For example, a head-to-head comparison between two index tests made within a study where all individuals receive both tests under evaluation will be regarded as more valid than comparisons made between studies where the two tests are evaluated in different studies (see Chapter 3). This chapter also provides guidance on assessing comparative accuracy studies (Yang 2021a).

8.2 Understanding bias and applicability

8.2.1 Bias and imprecision

Bias is a systematic error or deviation from the truth, either in the results of a study or in inference based on these results. Biases can act in either direction, leading to over-estimation or under-estimation of the true test accuracy.

It is difficult to evaluate whether the results of a single study are biased, and often impossible to predict the magnitude of a bias. However, when weaknesses or shortcomings are identified, judgements can be made of the risk of bias in an individual study, and sometimes its likely direction and size can be hypothesized.

Bias should not be confused with imprecision, which arises when an estimate is based on a small sample. Imprecision results from random error, whereas bias reflects a systematic error. Statistical analysis can appropriately describe the uncertainty in an estimate caused by random error, by using confidence intervals for example, but the confidence interval provides no information about the presence or absence of systematic error. In a systematic review we assess the risk of bias by examining the design and conduct of included studies.

8.2.2 Bias versus applicability

Traditionally the broader phrase 'assessment of methodological quality' was used (Moher 1996, Ioannidis 1998, Verhagen 2001), but in 2008 this was replaced by 'assessment of risk of bias' in Cochrane Reviews of interventions. Risk of bias focuses on whether the results of an individual study are valid and should be believed (Higgins 2008). Assessments focused on judging whether the methods used could introduce a risk of systematic error (i.e. bias) rather than on assessing the broader concept of methodological quality.

In addition to the assessment of risk of bias, a separate concept is considered in systematic reviews of test accuracy: the applicability of the results of individual studies to the review question (see Box 8.2.a). In test accuracy research the applicability of study findings to the review question matters, because estimates of accuracy can differ depending on population and setting characteristics (Ransohoff 1978, Mulherin 2002, Whiting 2013).

Box 8.2.a Bias and applicability

Bias: the degree to which estimates of diagnostic accuracy deviate from the truth. Risk of bias occurs if systematic flaws or limitations in the design or conduct of a test accuracy study produce study results that do not reflect the true accuracy of the test(s), as evaluated in the study.

Applicability: the extent to which findings from a primary study apply to the review question. Concerns regarding applicability may arise if the index test was evaluated in conditions that differ from those in the review question: if the index test was applied or interpreted differently, in a study group with different demographic or clinical features, or with a different definition of the target condition.

Concerns regarding applicability address partial or potential mismatches between the question addressed in an individual study and the question addressed in the systematic review. A study of a rapid antigen detection test in children, for example, may be done without serious limitations in design and conduct, and therefore be classified as at low risk of producing biased results. Yet if the review question addressed the performance of that same test in adults, the study findings may not directly answer the question of the systematic review, and review authors can express concerns regarding the applicability of the results of that study to their review.

A single diagnostic test can be applied in different ways and in different positions in the clinical pathway. For example, positron emission tomography–computed tomography (PET-CT) could be used as an add-on test after conventional staging to detect additional metastases in colorectal cancer patients with negative findings on earlier tests. In contrast, PET-CT could also be used as the first test in persons just diagnosed with colorectal cancer. Both are relevant uses of the test (potential review questions), but diagnostic accuracy may vary depending on the intended use of PET-CT. A study of PET-CT without methodological shortcomings, at low risk of producing biased results, may receive different applicability judgements, depending on the specific review question being addressed.

Poor and incomplete reporting of test accuracy studies frequently hampers the assessment of key features of design or conduct, making it difficult to judge the risk of bias or applicability (Reid 1995, Bossuyt 2003a, Lumbreras-Lacarra 2004, Smidt 2005, Korevaar 2015). When aspects of study design or execution are not reported, it is impossible for the reader to differentiate between poorly reported studies that used robust methodology, and studies applying poor methods that are likely to produce biased results. There are signs that the publication of the STARD statement (STAndards for the Reporting of Diagnostic accuracy studies) in 2003 (Bossuyt 2003a, Bossuyt 2003b) has improved the reporting of test accuracy studies, and it is hoped that this trend will continue (Korevaar 2014, Korevaar 2015). The STARD statement was updated in 2015 to STARD 2015 (Cohen 2016).

8.2.3 Biases in test accuracy studies: empirical evidence

The bias associated with particular study features can be examined in a field known as meta-epidemiology (Naylor 1997, Sterne 2002). A meta-epidemiological study starts with one or more sets of primary test accuracy studies. Each set examines a similar test accuracy question, but diagnostic tests and questions can differ across sets. Within each set, accuracy estimates of studies with and without a specific design feature are compared to get an empirical estimate of the bias associated with that feature. In the case of multiple sets of studies, estimates from each set can be included in meta-analysis, to obtain a more precise estimate of the bias.

While estimation and detection of the impact of certain design features have been well studied for randomized controlled trials in meta-epidemiology, only a few such projects have been undertaken for test accuracy studies (Lijmer 1999, Rutjes 2006). For many aspects of study methodology, conclusions about the existence or the magnitude of bias are based on case studies or on theoretical reasoning (Whiting 2013). Where meta-epidemiological evidence of bias for particular design items exists, we profile both empirical and theoretical evidence when discussing and illustrating specific issues within each domain.

8.3 QUADAS-2

8.3.1 Background

A large number of methodological quality assessment tools are available for test accuracy studies. A review of such tools in 2005 identified over 90 instruments (Whiting 2005). Because of the lack of a universal tool that can be used for test accuracy studies across all clinical domains and types of index tests, an initiative was started to develop a generic tool named QUADAS (Whiting 2003).

QUADAS-2 is a revision of the original QUADAS tool (Whiting 2011). QUADAS was formally revised and redesigned based on feedback from review authors, developments in the risk-of-bias tool for intervention reviews and new evidence about sources of bias and variation in test accuracy studies (Whiting 2013).

The QUADAS-2 tool has been widely accepted and is now the most frequently used tool for the assessment of risk of bias and applicability of primary studies in a systematic review of test accuracy; it is recommended by Cochrane for Cochrane Reviews of diagnostic test accuracy. The latest version of the tool, a template and background information can be found at www.quadas.org.

Key characteristics of the QUADAS-2 tool are:

- It evaluates the risk of bias of a study in four key domains:
    - participant selection (Section 8.4);
    - index test (Section 8.5);
    - reference standard (Section 8.6); and

  - flow and timing (Section 8.7).
- Signalling questions are included to facilitate judgements of the risk of bias.
- The first three domains are also assessed to identify concerns regarding applicability.
- For each domain, studies are rated as ‘low’, ‘high’ or ‘unclear’ for risk of bias; and ‘low’, ‘high’ or ‘unclear’ for concerns regarding applicability.

QUADAS-C, an extension of QUADAS-2 designed to assess risk of bias in primary studies comparing the accuracy of two or more tests, was published in 2021 (Yang 2021a). Later, we discuss QUADAS-2 as a starting point and explain where in the process the assessment should be modified for comparative accuracy studies, as recommended by QUADAS-C.

8.3.2 Risk-of-bias assessment

Signalling questions have been included to facilitate the process of reaching a judgement on the risk of bias for each domain. These questions are mainly factual questions that flag the potential for bias. Signalling questions are answered ‘Yes’, ‘No’ or ‘Unclear’, and are phrased such that ‘Yes’ indicates low risk of bias. If all signalling questions for a domain are answered ‘Yes’, then risk of bias for this domain can be judged as ‘low’. If any signalling question is answered ‘No’, this flags the potential for bias.

It is possible that one or more signalling questions are answered ‘No’ while the domain-level judgement is still ‘low risk of bias’. A ‘No’ answer should lead to a closer examination of the potential effect of design deficiencies, in terms of bias. If the shortcoming exists but is relatively minor, the final judgement could still be that the study is at low risk of producing biased results.

Review authors should rely on judgement to decide whether the potential source of bias flagged by the signalling question puts the study at risk of bias for that domain. The ‘unclear’ category should be used only when insufficient data are reported to permit a judgement.

When assessing risk of bias in a comparative accuracy study, the same structure and the same domains are assessed. The domains are first evaluated for each index test and then, with additional signalling questions, for the comparison. In Section 8.4 to Section 8.7, specific guidance for comparative accuracy will be discussed in addition to the guidance for single test accuracy.

8.3.3 Applicability assessment

Review authors are expected to record the information on which the judgement of applicability is based and then to rate their concern that the study does not match the review question. Concerns regarding applicability are judged as ‘low’, ‘high’ or ‘unclear’. Applicability judgements depend on precise formulations of review questions. Here also, the ‘unclear’ category should only be used when insufficient data are reported. The specific sections on each domain provide a more detailed explanation on how to judge the concerns regarding applicability in all relevant domains.

In comparative accuracy studies, concerns regarding applicability are captured by the QUADAS-2 assessments for each of the index tests in the comparison. Therefore, QUADAS-C does not include questions on applicability.

8.3.4 Using and tailoring QUADAS-2

Detailed information on how to answer the signalling questions and how to arrive at a risk-of-bias judgement for each domain are provided in Section 8.4. To make the process as transparent and objective as possible, review authors should produce guidance specifically for their review on how to answer each signalling question and how to use the answers to signalling questions to reach a judgement of risk of bias for each domain. Providing clear instructions on how to answer the signalling questions will improve consistency of interpretation between assessors. This should be done at the protocol stage, prior to conducting the QUADAS-2 assessment.

Review authors also have the option of adding specific signalling questions within a domain if particular issues relevant to their review are not fully covered by the standard QUADAS-2 tool. If review authors decide to add their own questions to the tool, they should phrase these in the same way as the questions in the main tool, to facilitate standardized reporting and interpretation of risk of bias and applicability in the review. Signalling questions should always be phrased so that a ‘Yes’ answer indicates no or limited risk of bias, and care should be taken that each question only covers one potential aspect of bias. For many reviews, the need to add questions will be limited.

8.3.5 Flow diagram

A flow diagram depicts the method of recruiting participants (for example, using a consecutive series of participants with specific symptoms suspected of having the target condition or separate groups of participants, one with and one without the target condition), the number of participants receiving which index test(s), which reference standard(s) and in which order. For a comparative accuracy study, drawing a flow diagram may help differentiate different study designs, e.g. (1) studies in which all eligible participants were given all index tests from (2) studies in which participants were given either one or both index tests, but of which only the fully paired subset was analysed. A good flow diagram will facilitate the answering of signalling questions and lead to better judgements of the risk of bias in a test accuracy study.

Review authors should draw a flow diagram for each included primary study if none is reported in the publication or if the published flow diagram is inadequate.

No single flow diagram structure can apply to all test accuracy designs; review authors will need to construct carefully the diagram for the study under evaluation. An example of a flow diagram, representing the recruitment of participants and the flow of participants, is given in Figure 8.3.a. Additional examples of flow diagrams can be found in the online supplementary material (8.S1 Example study flow diagrams).

8.3.6 Performing the QUADAS-2 assessment

Once the tool has been adapted to the review question and a final version with accompanying guidance has been produced, the next step is to develop a form to record the risk-of-bias and applicability assessment. On this form the review authors can record the answers to the signalling questions, and the judgement of risk of bias and applicability for each domain. This form should also contain space to record support for the judgement, a succinct summary of the stated facts given in the study report upon which the judgement is based. Review authors may find it useful to note where the description is taken from (e.g. the exact location within a report) for their own reference purposes.

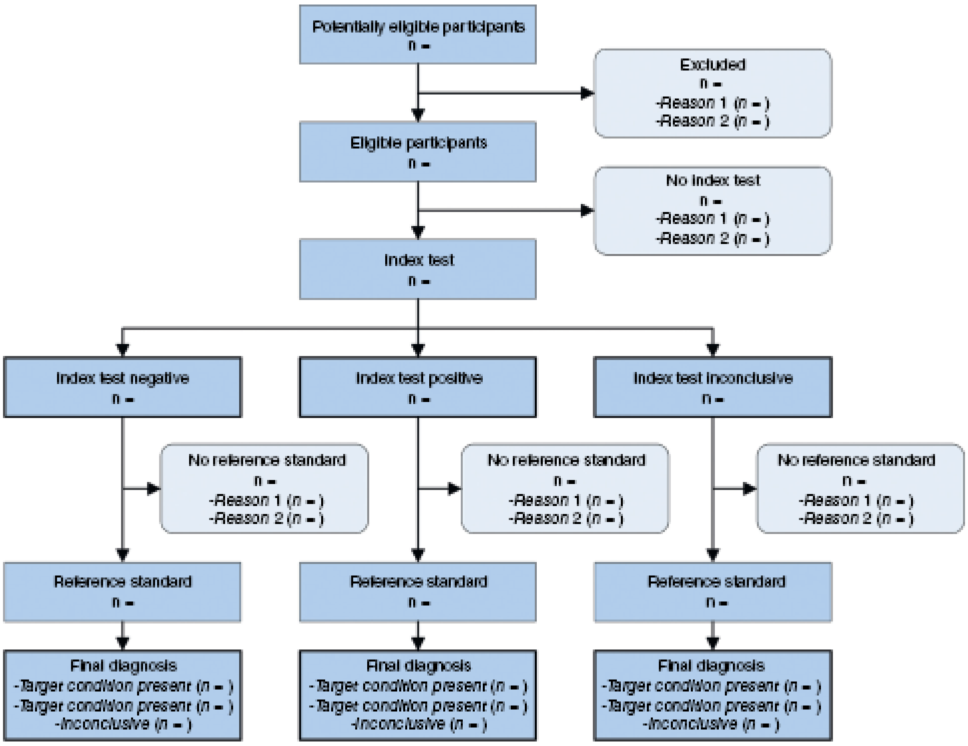


Figure 8.3.a Example of a study flow diagram for a test accuracy study

A period of testing and training is essential to improve the quality of forms and to calibrate answers between review authors, thereby lowering the number of future disagreements. It is usually a good idea for review authors to pilot the form on a diverse sample of at least five reports included in the review and then to meet and discuss any discrepancies in ratings.

Poor reporting is a major obstacle to risk-of-bias assessment because it obscures whether a design feature was correctly applied but not reported, or was inappropriately applied, potentially introducing bias. Contacting study authors for further information is one method of dealing with poor reporting (Pai 2003) (see Chapter 7, Section 7.2.2). This process may provide useful information and so is a worthwhile undertaking. To reduce the risk of overly positive answers, caution should be exercised in ensuring that study authors are not asked leading questions. For example, it is better to ask the study authors to describe their processes, such as ‘How did the physicians in the study decide whether an individual underwent biopsy?’ or ‘What information was the radiologist given about a patient?’, rather than directly asking them to make a judgement as to whether their study was at risk of producing biased results.

Preferably, at least two review authors should perform the assessment independently. The review authors should have relevant knowledge of both the methodological issues of test accuracy studies and the clinical topic area. There should be an explicit procedure for resolving disagreements among review authors. This may involve discussion between review authors or referral to a third author if agreement cannot be reached.

Blinding for journal, and/or names of study authors, is not recommended, as it can be difficult to achieve in practice, is time consuming and its benefit is uncertain (Jadad 1996, Berlin 1997, Kjaergard 2001, Gluud 2008).

Review authors are encouraged to clarify, for each question, why they assigned a judgement for the corresponding domain, in light of potential problems being present.

8.4 Domain 1: Participant selection

The first domain relates to participant selection: how study participants were identified, contacted and included in the study, and whether this could have introduced bias. Applicability refers to the match, or the lack thereof, between study participants and the target population, as defined in the review question.

8.4.1 Participant selection: risk-of-bias signalling questions (QUADAS-2)

The key issue to address here is whether the selection of participants could have introduced bias. QUADAS-2 includes three signalling questions to flag the risk of bias for the participant selection domain.

Signalling question 1: Was a consecutive or random sample of participants enrolled?

A test accuracy study should ideally enrol a consecutive series of participants suspected of having the target condition, or a random sample thereof. Studies that recruit participants in a different way may produce estimates of test accuracy that do not reflect the performance of the test in clinical practice.

Enrolling a consecutive series of participants, or a random sample thereof, ensures that study participants are likely to be representative of those undergoing testing in practice. It guarantees that persons more difficult to diagnose are not purposefully excluded from a study, or that persons for whom there is a strong suspicion of the target condition are not preferentially selected for inclusion in the study. Review authors should note that the term ‘consecutive’ is used liberally by authors in primary study reports, and does not always indicate a truly consecutive series of participants suspected of having the target condition.

A large number of studies reporting on the accuracy of a test have included a study group that does not form a consecutive series of participants suspected of having the target condition. Instead, they have recruited what can best be described as convenience samples:

they used results at hand or recruited study participants in a non-systematic way, for example.

Examples are studies that searched hospital records for patients who, in the past, had undergone both the index test and the reference standard, or comparative accuracy studies enrolling only participants in whom a second index test was deemed necessary. Such situations should be flagged as having non-consecutive series (signalling question), but when judging the risk of bias for this domain, review authors need to consider whether the sampling strategy is likely to have introduced bias.

Critical in this judgement is whether the included group of participants can still be considered representative of the target population. This judgement can be difficult, as many studies do not explicitly report on the processes for identifying, selecting and inviting potentially eligible study participants.

Signalling question 2: Was a case-control design avoided?

Some studies use one selection process to recruit participants already known to have the target condition and a different one to recruit participants known not to have the target condition (see ‘multiple groups’ in Chapter 3) (Rutjes 2005). This used to be referred to as a ‘case-control design’. The selection of ‘cases’ (those with the target condition) and ‘controls’ (those without the target condition) then becomes a critical issue, as estimates of sensitivity and specificity will be biased if cases and/or controls systematically differ from individuals with and without the target condition within the target population.

A clear example is a study including one group of participants with the target condition and a second group of healthy controls (Lijmer 1999, Pai 2003). Including healthy, asymptomatic controls, such as blood donors, can be expected to lead to fewer false positive test results compared to the inclusion of participants with symptoms suggesting the presence of the target condition. As a consequence, estimates of specificity are likely to be inflated.

Accuracy studies with multiple groups of participants, each recruited in a different way, often also include more typical or more extreme phenotypes of the target condition. These typical cases are more likely to receive true-positive test results, compared to a group of participants that represents the full spectrum of the target condition, from less severe to more severe.

A design in which those with the target condition and those without are each randomly sampled from one consecutive series of participants suspected of having the target condition will not lead to biased estimates (Biesheuvel 2008, Pepe 2008). Such a study bears similarities with what is known as a nested case-control study in aetiological research. This is an example of a situation in which a signalling question will be answered with ‘No’ (red flag), without assigning a high risk of bias for this domain.

Signalling question 3: Did the study avoid inappropriate selection criteria?

This signalling question draws attention to whether a study excluded persons at the point of enrolment, even though such persons are or will be evaluated with the index test in clinical practice. This may influence estimates of test accuracy for that study, which then no longer reflects the accuracy of the test in clinical practice.

This question about inappropriate selection criteria relates only to the eligibility criteria and the recruitment process, i.e. how participants were invited for the study; exclusion of study participants after enrolment is covered in the flow and timing domain (domain 4).

Examples include the exclusion of patients with high body mass index (BMI) in ultrasound studies for abdominal pain, or the exclusion of patients with pre-existing lung disease in studies examining the accuracy of CT in patients with suspected pulmonary embolism. In both situations, patients who potentially were more difficult to diagnose were not enrolled in the study, which may lead to inflated estimates of diagnostic accuracy.

It is possible that a study did not have explicit exclusion criteria, but specific subgroups undergoing the test in practice are notably absent from the study group. This can sometimes be deduced from a table with baseline characteristics. In that case, study-specific accuracy estimates may also be biased, if the risk of false positive test results, or the risk of false negative test results, is known to vary depending on patient characteristics. For example, the specificity of D-dimer, a test for pulmonary embolism, is known to be lower in higher age groups. A study to estimate its accuracy in an emergency department may produce biased estimates if that study failed to include elderly patients.

Given the complexity of the problem underlying this signalling question, review authors are encouraged to generate detailed guidance on how to answer this signalling question for their review.

8.4.2 Participant selection: additional signalling questions for comparative accuracy studies (QUADAS-C)

Informative comparisons of the accuracy of tests should reflect the performance of these tests in the same target population. If one test is evaluated in one group of participants and the other tests in different groups, any difference in accuracy may be due, in part or in full, to differences between these groups, rather than to the tests evaluated. QUADAS-C contains four additional signalling questions for the participant selection domain to cover specific sources of bias that are relevant for comparative accuracy studies.

Additional signalling question 1: Was the risk of bias for each index test judged 'low' for this domain?

We first consider whether the risk-of-bias judgement in QUADAS-2 for the participant selection domain was judged as 'low' for all index tests in the comparison. If not, the study group or combination of study groups may not be representative of the target population. This may be the case, for example, if participants difficult to evaluate with one of the index tests were excluded.

This question should be answered with 'No' if at least one of the index tests was judged to be at 'unclear' or 'high' risk of bias for this domain.

Additional signalling question 2: Was a fully paired or randomized design used?

This question addresses whether an appropriate design was used to compare multiple index tests. The preferred option for comparing multiple tests is to perform all index tests in all participants (fully paired design); an alternative is random allocation to one index test or to another (randomized design).

If neither a fully paired nor randomized design was used, those receiving one index test and those receiving another may not be comparable in terms of factors that affect test accuracy. An example is allocation of study participants to one specific index test based on presenting features, and allocation of all other participants, without such features, to a different index test.

It may be challenging to distinguish a fully paired design with missing index test results from a 'partially paired' design, in which some participants receive all index tests while others receive only one of the tests. Examining the study protocol (if available) or careful examination of the Methods section of the study report may clarify if this partially paired design was the design intended by the study investigators (Yang 2021b).

Additional signalling question 3: Was the allocation sequence random? (only applicable to randomized designs)

The allocation sequence in a randomized comparative accuracy study should be generated by a truly random process. This could be achieved by computer-based random number generators, drawing lots, random number tables, or other methods that involve a truly random component. Assigning tests to a person based on admission date or date of birth is not a fully random process. Details about the randomization process may often be poorly reported in studies and it may be necessary to contact the study authors for further details.

Additional signalling question 4: Was the allocation sequence concealed until participants were enrolled and assigned to index tests? (only applicable to randomized designs)

A failure to conceal the allocation when deciding about the eligibility of a potential study participant may lead to selective enrolment of participants. If so, the groups assigned to the respective tests may no longer be comparable and estimates of relative accuracy may be biased.

Concealment of allocation may be achieved, for example, by assigning the randomly generated number only after the participant was enrolled, by allocating participants by means of a central randomization procedure, or by keeping random numbers in sequentially numbered opaque and sealed envelopes.

Table 8.4.a summarizes the signalling questions used in the risk-of-bias assessment for the participant selection domain. Additional signalling questions are listed in Table 8.4.b.

Table 8.4.a Summary of risk-of-bias assessment for Domain 1: Participant selection

Signalling question 1: Was a consecutive or random sample of participants enrolled?	
Yes	If the method of recruitment was consecutive or a random sample was taken from a consecutive series
No	If there is evidence for non-consecutive or non-random inclusion of eligible participants
Unclear	If the method of participant sampling is unclear
Signalling question 2: Was a case-control design avoided?	
Yes	If a single group of participants suspected of having the target condition was recruited
No	If two or more groups of participants were separately recruited, with one group consisting of participants already known to have the target condition, and a second group consisting of healthy controls, or a group of participants with a specific alternative condition
Unclear	If the selection of participants is unclear
Signalling question 3: Did the study avoid inappropriate selection criteria?	
Yes	If all patients who would undergo the test in practice were eligible for the study
No	If selection criteria were defined such that specific subgroups within the target population were ineligible, or if such patients were not invited to the study
Unclear	If it is unclear whether inappropriate selection criteria were used
Risk-of-bias assessment: Could the selection of participants have introduced bias?	
Low	If the answer to all signalling questions is 'Yes' then risk of bias can be considered low
High	If the answer to any of the signalling questions is 'No', there is a potential for bias. If one or more of the answers is 'No', the judgement could still be low risk of bias, but specific reasons why the risk of bias can be considered low should be provided

Unclear	If relevant information is missing for all or some of the signalling questions, and none of the answers to signalling questions is judged to put the study at high risk of bias

Table 8.4.b Comparative accuracy, additional signalling questions for the participant selection domain (QUADAS-C)

<table>
<tr><td colspan="2"><b>Additional question 1: Was risk of bias for this domain judged ‘low’ for all index tests?</b></td></tr>
<tr><td>Yes</td><td>if risk of bias in QUADAS-2 was judged ‘low’ for all index tests</td></tr>
<tr><td>No</td><td>if risk of bias in QUADAS-2 was judged to be ‘high’ or ‘unclear’ for one or more index tests</td></tr>
<tr><td colspan="2"><b>Additional question 2: Was a fully paired or randomized design used?</b></td></tr>
<tr><td>Yes</td><td>If it is clear from the study report or from the study protocol that, by design, all participants would receive all tests (fully paired design) or that participants would be randomly allocated to one of the tests (randomized design)</td></tr>
<tr><td>No</td><td>Neither a fully paired nor randomized design was used to allocate participants to index tests</td></tr>
<tr><td>Unclear</td><td>If it was unclear how participants were allocated to the tests being compared</td></tr>
<tr><td colspan="2"><b><i>Additional question 3: Was the allocation sequence random? (only applicable to randomized designs)</i></b></td></tr>
<tr><td>Yes</td><td>If authors used a truly random process to generate the allocation sequence</td></tr>
<tr><td>No</td><td>If authors generated an allocation sequence that was not the result of a truly random process</td></tr>
<tr><td>Unclear</td><td>If authors only stated that the study was randomized without further information</td></tr>
<tr><td colspan="2"><b>Additional question 4: Was the allocation sequence concealed until participants were enrolled and assigned to index tests? (only applicable to randomized designs)</b></td></tr>
<tr><td>Yes</td><td>If authors report an appropriate method to conceal allocation</td></tr>
<tr><td>No</td><td>If authors did not conceal the allocation at the time of enrolment</td></tr>
</table>

Unclear	If it is unclear whether allocation was concealed
Risk-of-bias judgement:	
Low	If the answer to all additional questions is 'Yes', then risk of bias can be considered low
High	If the answer to any of the signalling questions is 'No', there is a potential for bias. If one or more of the answers is 'No', the judgement could still be low risk of bias, but specific reasons why the risk of bias can be considered low should be provided
Unclear	If relevant information is missing for all or some of the signalling questions, and none of the answers to signalling questions is judged to put the study at high risk of bias

8.4.3 Participant selection: concerns regarding applicability

The question to be answered is whether there are concerns that the included participants and setting do not match the review question. It is possible that the primary study was designed to answer a research question that differs from the review question. For that reason, the group of participants in a study may not match the population that is targeted by the review question.

Concerns regarding applicability may arise if study participants differ from those targeted by the review question in terms of severity of the target condition, demographic features, presence of alternative conditions or comorbidity, setting of the study, or previous tests. For example, larger tumours are more easily detected on imaging than smaller ones, and larger myocardial infarctions lead to higher levels of cardiac enzymes than small infarctions, making them easier to detect and so increasing estimates of sensitivity. A D-dimer test may be evaluated in an emergency department setting, while the review addresses the question of its accuracy when performed by general practitioners.

If a test accuracy study recruits one group of patients with known severe disease and a second group of healthy controls, it will be flagged here, but also assessed as being at high risk of bias in the participant selection domain, because such a study will produce inflated estimates of sensitivity and specificity.

Table 8.4.c Summary of main issues when evaluating applicability for the participant selection domain

Domain 1: Participant selection	**Concerns regarding applicability**

Are there concerns that the included participants and setting do not match the review question?	
Low	If the spectrum of participants (inclusion and exclusion criteria, setting, prior testing) matches the pre-stated requirements in the review question
High	If the spectrum of participants does not fully match the pre-stated requirements in the review question
Unclear	If there is insufficient information available to make a judgement about applicability

Table 8.4.c summarizes the main issues when evaluating applicability for the participant selection domain.

8.5 Domain 2: Index test

This domain relates to the index test: whether conduct or interpretation of the index test could have introduced bias, and whether the index test, as used in the study, reflects the one in the review question (applicability).

8.5.1 Index test: risk-of-bias signalling questions (QUADAS-2)

The key issue to answer is whether the conduct or interpretation of the index test could have introduced bias. Two signalling questions are included to flag potential issues within a study that could lead to bias.

Signalling question 1: Were the index test results interpreted without knowledge of the results of the reference standard?

Interpretation of index test results may be influenced by knowledge of the reference standard, thereby artificially increasing concordance between them, leading to higher estimates of accuracy (Whiting 2013). The potential for bias is related to the degree of subjectivity involved in interpreting the index test. When an index test requires an interpretation, interpreters are more likely to be influenced by the results of the reference standard than with a fully automated or objective test. It therefore matters whether the unblinded interpretation of the index test could have led to bias.

Whether or not blinding was undertaken in a study may not be stated explicitly in the study report or protocol; if index tests and reference standard are undertaken and interpreted in a specified order, the first test will have been performed without access to the results of the second.

Where the index test and reference standard were undertaken by different individuals, a degree of ambiguity may exist about what information was available for each test. In some instances, knowledge of standard laboratory practices may allow reasonable assumptions to be made, but confirmation from the study authors is always desirable.

Signalling question 2: If a threshold was used, was it pre-specified?

For index tests that give ordered categories, counts or continuous measurements, a threshold is applied to classify test results as positive or negative (see Chapter 4). Selecting a positivity threshold in a study during data analysis, instead of during study design or protocol development, can lead to inflated estimates of test accuracy. This is especially the case when the selection is based on calculations of accuracy. In many studies, for example, authors select the positivity threshold that maximizes the sum of sensitivity and specificity (Youden index, see Chapter 4) in the data they have collected.

The reason post-hoc selection of the positivity threshold causes bias is that the observed ROC curve will fluctuate considerably around the true underlying ROC curve, especially in small studies. A data-driven selection of a threshold is likely to result in selecting a point estimate that is by chance further away from the true underlying ROC curve. Subsequently using that same threshold in an independent sample of patients is likely to lead to lower estimates of accuracy (Leeflang 2008). Pre-specifying the threshold before collecting or analysing will prevent this bias.

The following should be considered when judging whether a data-driven threshold is likely to produce a biased result. First, the magnitude of the bias is inversely related to sample size (Leeflang 2008). In a large study the observed ROC curve will be closer to the true underlying curve. Bias due to data-driven selection of thresholds is therefore smaller in studies with at least 100 participants with and, at the same time, 100 without the target condition (Leeflang 2008). Second, researchers may have used more robust methods for selecting a threshold, for example by first fitting smooth ROC curves based on parametric or non-parametric assumptions. Such methods generally reduce bias, but their performance depends on whether underlying assumptions are met (Leeflang 2008).

Ideally, study authors should use a positivity threshold that was predefined in the study protocol, or one that is specified by the manufacturer of the test. In the case of truly dichotomous index test results, this signalling question can be answered ‘Yes’, because there is no risk of bias due to selection of thresholds.

8.5.2 Index test: additional signalling questions for comparative accuracy studies (QUADAS-C)

For comparative accuracy studies, the signalling questions should first be answered for each index test separately. There are four additional questions to consider for the index test domain when dealing with comparative accuracy studies.

Additional signalling question 1: Was the risk of bias for each index test judged ‘low’ for this domain?

If the risk-of-bias judgement in QUADAS-2 for the index test domain was judged as ‘high’ for one or more index tests, then the comparison between these index tests would likely be biased as well. For example, if index test A was interpreted with knowledge of the results of the reference standard, then not only is the estimated accuracy of A possibly biased, but also the estimate of the difference in accuracy between A and another index test B. If estimates of both test A and test B are judged to be at high risk of bias, the comparison between them will still be at high risk of bias, as the direction and magnitude of bias affecting each index test may not be the same.

Additional signalling question 2: Were the index test results interpreted without knowledge of the results of the other index test(s)? (only applicable if participants received multiple index tests)

Interpreting an index test while knowing the results of another may artificially increase concordance between the results of these index tests and could therefore bias the comparison. Such a bias could be mitigated by blinding the index test interpreter from the results of the other index test(s).

When answering this signalling question, similar considerations from the QUADAS-2 question 'Were the index test results interpreted without knowledge of the results of the reference standard?' apply: the potential for bias depends on the degree of subjectivity involved in test interpretation. The order in which the tests were performed and interpreted may suggest whether or not the interpretation was blinded.

In addition, authors of systematic reviews comparing the accuracy of one index test versus a combination of that test with another should note that blinding is not always necessary. For example, in a comparative accuracy study of ultrasound versus ultrasound followed by CT, CT readers do not need to be blinded to the results of ultrasound if ultrasound results are also available to the CT interpreter in practice. In this example, only the ultrasound interpreter needs to be blinded to the results of CT.

Additional signalling question 3: Is undergoing one index test unlikely to affect the performance of the other index test(s)? (only applicable if participants received multiple index tests)

It is possible for an index test to influence or interfere with the performance of subsequent index tests. Possible reasons for interference include patient or investigator fatigue (if performing the index tests requires mental or physical effort) and contrast agents or relaxants used for the first imaging test that may also affect the performance of the second test. If so, accuracy estimates for the second index test may not reflect its accuracy when performed as a first test.

Randomizing the order of index tests is sometimes done with the intention of preventing this bias, but a random order is expected to prevent bias only under the following circumstances: (1) the first test interferes with the performance of the second test and the other way round, if the order is reversed; and (2) the bias affecting each test has the same direction and magnitude on the scale of interest (absolute or relative difference in accuracy).

Additional signalling question 4: Were the index tests conducted and interpreted without advantaging one of the tests?

There may be ways of handling or interpreting the index tests that may advantage one test over the other. An example is when one index test is interpreted by an experienced reader whereas the competing index test is interpreted by a less experienced reader. One index test may be performed in a specialized clinic whereas the other is performed in a busy general practice. One index test may be performed on fresh samples while frozen samples are used for the other test. Differences between the index tests that reflect differences in clinical practice should not be considered to be a source of bias.

Table 8.5.a summarizes the signalling questions and how these should be answered to assess the risk of bias for the index test domain. Additional signalling questions are listed in Table 8.5.b.

Table 8.5.a Summary of risk-of-bias assessment for Domain 2: Index test

Signalling question 1: Were the index test results interpreted without knowledge of the results of the reference standard?	
Yes	If the index tests were interpreted without access to the results of the reference standard (explicitly stated or assumed from typical practice)
No	If the index tests were clearly interpreted with access to the results of the reference standard
Unclear	If it is unclear whether index tests were interpreted without access to the results of the reference standard
Signalling question 2: If a threshold was used, was it pre-specified?	
Yes	If the positivity threshold was pre-specified during protocol development
No	If the positivity threshold was based on an analysis of data collected in the study itself
Unclear	If it is unclear how the positivity threshold for the index test was selected
Risk-of-bias judgement: Could the conduct or interpretation of the index test have introduced bias?	
Low	If the answer to all signalling questions is ‘Yes’, then risk of bias can be considered low
High	If the answer to any of the signalling questions is ‘No’, there is a potential for bias. If one or more of the answers is ‘No’, one could still assign a low risk of

	bias but specific reasons for doing so should be provided. For example, if information on reference standard results was available to the person interpreting the index test, but the index test results were generated by a device, then this is unlikely to have introduced bias
Unclear	If relevant information is missing for all or some of the signalling questions, and none of the answers to signalling questions is judged to put the study at high risk of bias

Table 8.5.b Comparative accuracy, additional signalling questions for the index test domain (QUADAS-C)

<table>
<tr><td colspan="2"><b>Additional question 1: Was the risk of bias for each index test judged ‘low’ for this domain?</b></td></tr>
<tr><td>Yes</td><td>If risk of bias in QUADAS-2 was judged ‘low’ for both or all index tests</td></tr>
<tr><td>No</td><td>If risk of bias in QUADAS-2 was judged to be ‘high’ or ‘unclear’ for one or more index tests</td></tr>
<tr><td colspan="2"><b>Additional question 2: Were the index test results interpreted without knowledge of the results of the other index test(s)? (only applicable if participants received multiple index tests)</b></td></tr>
<tr><td>Yes</td><td>If it is likely that the interpretation of each index test was done without knowledge of the results of the other index test(s) in the study</td></tr>
<tr><td>No</td><td>If it is likely that the interpreter of any one of the index tests was aware of the results of any of the other index tests</td></tr>
<tr><td>Unclear</td><td>If it is unclear whether the interpreters of the index tests were aware of the results of the other index tests</td></tr>
<tr><td colspan="2"><b>Additional question 3: Is undergoing one index test unlikely to affect the performance of the other index test(s)? (only applicable if participants received multiple index tests)</b></td></tr>
<tr><td>Yes</td><td>If the conduct of the index tests could not have influenced the performance of any of the other index tests</td></tr>
<tr><td>No</td><td>If it is likely that the conduct of one index test influenced the performance of subsequent index test(s)</td></tr>
</table>

Unclear	If it is unclear whether an index test could have influenced the performance of subsequent index test(s)
Additional question 4: Were the index tests conducted and interpreted without advantaging one of the tests?	
Yes	If all index tests were performed and interpreted under similar circumstances
No	If there are substantial differences in how the index tests were performed or interpreted, and these differences are likely to affect the accuracy of one of the tests more favourably relative to the other test(s)
Unclear	If the study provides insufficient information to judge whether the index tests were done under similar circumstances
Risk-of-bias assessment: Could the conduct or interpretation of the index tests have introduced bias in the comparison?	
Low	If the answer to all additional questions is 'Yes', then risk of bias can be considered low
High	If the answer to any of the signalling questions is 'No', there is a potential for bias. If one or more of the answers is 'No', the judgement could still be low risk of bias, but specific reasons why the risk of bias can be considered low should be provided
Unclear	If relevant information is missing for all or some of the signalling questions, and none of the answers to signalling questions is judged to put the study at high risk of bias

Table 8.5.c Summary of the main issues when evaluating applicability for the index test domain

Domain 2: Index test	**Concerns regarding applicability**
Are there concerns that the index test, its conduct or its interpretation differ from the review question?	
Low	If the index test technology and the way the test has been applied and interpreted in the study match the pre-stated requirements in the review question

Domain 2: Index test	**Concerns regarding applicability**
High	If there are differences in index test technology, execution and interpretation between the study and the review question
Unclear	If there is insufficient information available to make a judgement about applicability for this domain

8.5.3 Index test: concerns regarding applicability

Variations in test technology, execution or interpretation may affect test accuracy. If index test methods vary from those specified in the review question, there may be concerns regarding applicability (Stengel 2005). An accuracy study of an imaging modality, for example, may use a consensus read by three radiologists, whereas in clinical practice such images would be read by only one person.

Table 8.5.c summarizes the main issues when evaluating applicability for the index test domain.

8.6 Domain 3: Reference standard

The central issue in this domain is whether the reference standard, its conduct or interpretation could have introduced bias, and whether the target condition, as detected by the reference standard, reflects the review question.

8.6.1 Reference standard: risk-of-bias signalling questions (QUADAS-2)

Two signalling questions in QUADAS-2 flag potential issues within a study that could lead to bias.

Signalling question 1: Is the reference standard likely to correctly classify the target condition?

Measures of test accuracy are calculated by assuming that the reference standard is error free; any disagreement between the index test and the reference standard results will be classified as either a false positive or false negative index test result.

Bias in estimates of accuracy is likely if the reference standard does not correctly classify study participants. If there are misclassifications by the reference standard, an index test result that is classified as a false positive may in fact be a true positive, and a false negative index test result may be a true negative. As a consequence, imperfect reference standards can bias estimates of the accuracy of an index test (Boyko 1988).

The net effect of misclassification by the reference standard can be an upward or downward bias in estimates of test accuracy, depending on the frequency of the misclassification and the type of misclassification. Under-estimation can occur when the index test and the reference standard measure different aspects of the target condition, such that errors in the reference standard are unrelated to errors in the index test. Accuracy may be overestimated when the index test and reference standard measure similar aspects of the target condition, such that errors in the reference standard are likely to occur together with errors in the index test (van Rijkom 1995, Biesheuvel 2007).

Knowledge about imperfections in the reference standard cannot be obtained from a study in which the accuracy of one or more index tests is evaluated. Evidence about the likelihood of reference standard misclassifications will have to be found in other studies or systematic reviews of such studies, for example studies that assessed the reproducibility and repeatability of the reference standard. An example are tandem colonoscopy studies, in which consenting participants undergo two same-day colonoscopies with polypectomy (Zhao 2019). Another example is a study in which liver biopsies are evaluated by not one but a team of hepatopathologists, who are invited to read these biopsies independently, blinded from one another's judgement (Davison 2020).

A perfect reference standard – a gold standard – will rarely be available; most reference standards have imperfections. Some degree of imperfection may still be acceptable, if the frequency of misclassification errors from the reference standard is low. At the protocol stage, review authors will define the preferred reference standard. Many reviews will then restrict inclusion of studies based on the acceptability of the reference standard in light of the target condition. In addition, the criteria for scoring this signalling question as 'Yes' for included studies will be defined at the protocol stage.

If one expects the index test to outperform the reference standard, assessing the accuracy of such a test by assessing index test results against this reference standard will not be helpful (see Chapter 3). Reviews of test accuracy should not be undertaken in these circumstances without careful consideration of the methodological issues (Glasziou 2008, Reitsma 2009).

Signalling question 2: Were the reference standard results interpreted without knowledge of the results of the index test?

The results of the reference standard are ideally obtained without knowledge of the results of the index test(s). Knowledge of the index test results can influence the interpretation of the reference standard (Whiting 2013). The danger is that this may artificially increase agreement between the index test and the reference standard results, leading to inflated estimates of test accuracy. This is also known as test-review bias (Ransohoff 1978).

The potential for bias is related to the degree of subjectivity involved in interpreting the results of the reference standard. When a reference standard requires a more subjective reading, interpreters are more likely to be influenced by the results of the index test than for a fully automated reference standard or a reference standard with explicit criteria.

An example is the use of an expert panel as the reference standard, where the expert panel has access to the index test results (Bertens 2013). In a study examining the accuracy of magnetic resonance imaging (MRI) in detecting multiple sclerosis, the reference standard was a panel-based assignment of a final diagnosis, based on all available information, including MRI results, cerebrospinal fluid analysis and clinical follow-up of participants. To avoid bias, studies using an expert panel may decide not to present the results of the index test to the panel, to prevent giving these results too much weight when deciding whether the target condition is present or not.

An extreme form occurs if index test results are a formal component of the reference standard; i.e. when the result of the index test is incorporated into the evidence used to conclude that the target condition is present or absent. The resulting bias is known as incorporation bias (Ransohoff 1978, Worster 2008).

8.6.2 Reference standard: additional signalling questions for comparative accuracy studies (QUADAS-C)

For comparative accuracy studies, the signalling questions should first be answered for each index test separately. There are two additional questions to consider for the reference standard domain when dealing with comparative accuracy studies.

Additional signalling question 1: Was the risk of bias for each index test judged ‘low’ for this domain?

Similar to the previous domains, we first ask whether the risk of bias in QUADAS-2 for the reference standard domain was judged as ‘low’ for the index tests being compared. If that is not the case, the comparison between these index tests could be biased. The use of an imperfect reference standard in all study participants would not only affect accuracy estimates for each index test, but could also bias estimates of comparative accuracy.

Table 8.6.a Summary of risk-of-bias assessment for Domain 3: Reference standard

Signalling question 1: Is the reference standard likely to correctly classify the target condition?	
Yes	If a reference standard has been used that is considered by clinical experts to be error free for the target condition
No	If a reference standard has been used that is known to lead to misclassifications
Unclear	If it is unclear exactly what reference standard was used
Signalling question 2: Were the reference standard results interpreted without knowledge of the results of the index test?	
Yes	If it is clear that the index test results were not available to those interpreting the reference standard results
No	If it is clear that the index test results were available to those interpreting the reference standard results
Unclear	If it is unclear whether the results of the index test were available to those interpreting the reference standard results
Risk-of-bias judgement: Could the reference standard, its conduct or its interpretation have introduced bias?	
Low	If the answer to all signalling questions is ‘Yes’, then risk of bias can be considered low
High	If the answer to any of the signalling questions is ‘No’, there is a potential for bias. If one or more of the answers is ‘No’, the judgement could still be low risk of bias, but specific reasons why the risk of bias can be considered low should be provided
Unclear	If relevant information is missing for all or some of the signalling questions, and none of the answers to signalling questions is judged to put the study at high risk of bias

Additional signalling question 2: Did the reference standard avoid incorporating any of the index tests?

The second signalling question in this domain is whether one or more index tests were part of the reference standard, i.e. used to decide on the presence or absence of the target condition. If this was the case for one index test (A) and not for another index test (B), the accuracy of A will be over-estimated, leading to a biased estimate of the difference in accuracy between the index tests. If both index tests A and B are part of the reference standard, there is still risk of a biased comparison, as the two index tests may not contribute equally to the final diagnosis.

Table 8.6.a summarizes the signalling questions and how these should be answered to assess the risk of bias for the reference standard domain. Additional signalling questions are listed in Table 8.6.b.

8.6.3 Reference standard: concerns regarding applicability

The question to be answered is whether there are concerns that the target condition as detected by the reference standard does not match the review question.

Many diseases and conditions are not dichotomous; instead they present as a spectrum, ranging from less to more severe. When specifying the review question(s), review authors should be specific about the definition of the target condition in their review, and should reflect on what reference standard is best able to detect that target condition. The reference standard should then be selected in light of this target condition. Different reference standards may detect different forms of the target condition.

Table 8.6.b Comparative accuracy, additional signalling questions for the reference standard domain (QUADAS-C)

Additional question 1: Was the risk of bias for each index test judged ‘low’ for this domain?	
Yes	if risk of bias in QUADAS-2 was judged ‘low’ for both or all index tests
No	if risk of bias in QUADAS-2 was judged to be ‘high’ or ‘unclear’ for one or more index tests
Additional question 2: Did the reference standard avoid incorporating any of the index tests?	
Yes	If none of the index tests was part of the reference standard
No	If one or more of the index tests were used to decide on presence or absence of the target condition

Unclear	If it is unclear whether any of the index tests were part of the reference standard
Risk-of-bias assessment: Could the reference standard, its conduct or its interpretation have introduced bias in the comparison?	
Low risk of bias	If the answer to all additional questions is 'Yes', then risk of bias can be considered low
High risk of bias	If the answer to any of the signalling questions is 'No', there is a potential for bias. If one or more of the answers is 'No', the judgement could still be low risk of bias, but specific reasons why the risk of bias can be considered low should be provided
Unclear risk of bias	If relevant information is missing for all or some of the signalling questions, and none of the answers to signalling questions is judged to put the study at high risk of bias

Table 8.6.c Summary of the main issues when judging applicability for the reference standard domain

Domain 3: Reference standard	**Concerns regarding applicability**
Are there concerns that the target condition as detected by the reference standard does not match the review question?	
Low	If the reference standard, as used in the study, detects the target condition defined in the review question
High	If the reference standard, as used in the study, does not detect the same (form of) target condition as defined in the review question
Unclear	If there is insufficient information available to make a judgement about applicability for this domain

For example, when defining urinary tract infection, the reference standard is generally based on specimen culture, but the threshold above which a result is considered positive may vary (Tullus 2019). In studies evaluating screening tests for colorectal cancer, colonoscopy is the accepted reference standard, but the definition of what the test aims to detect (referred to as 'advanced neoplasia') can vary between studies. Table 8.6.c summarizes the main issues when judging applicability for the reference standard domain.

8.7 Domain 4: Flow and timing

The fourth and final domain focuses on the flow of participants through a test accuracy study and the timing of the index test(s) and the reference standard. This domain is only assessed in terms of consequences for risk of bias. The central issue is whether the participant flow could have introduced bias.

In the ideal test accuracy study, each participant receives both the index test and the reference standard at the same point in time. Bias may be introduced if the time interval between the index test and the reference standard is not appropriate. Bias may also be introduced if not all study participants are included in the analysis.

8.7.1 Flow and timing: risk-of-bias signalling questions (QUADAS-2)

QUADAS-2 includes four signalling questions to flag potential issues within a study that could lead to bias in this domain.

Signalling question 1: Was there an appropriate interval between the index test and reference standard?

Ideally, the index test and reference standard are performed in the same participant at the same time. If there is a delay, natural disease progression and/or treatment between index test and reference standard can lead to changes in the target condition: it may be present at the time of testing with the index test, but resolved by the time the reference standard is performed, or the other way around. If the target condition resolved during the time interval, a true positive index test result may be erroneously classified as a false positive. A delay is therefore a potential cause of bias. The potential for changes in the target condition because of a delay, and the potential for bias, will vary between conditions.

Conversely, when the reference standard is based on observations during follow-up, a minimum length may be required to capture the symptoms or signs indicating that the target condition was present when the index test was performed. For example, for the evaluation of MRI for the early diagnosis of multiple sclerosis, a minimum follow-up period of around 10 years is required to be confident that all participants who will go on to fulfil the diagnostic criteria for multiple sclerosis will have done so (Whiting 2006). If there are concerns that (effective) treatments may have been applied during the time interval, this potential for bias may be highlighted in the flow and timing domain.

Starting treatment for the target condition before the reference standard has been performed (i.e. before the end of follow-up) can affect results, since it may lead to recovery. When judging the potential for bias, it is important to estimate the proportion of participants in whom the interval was inappropriate. The magnitude of the bias will increase if the number of participants outside the appropriate interval is larger.

The acceptable length of the interval between index test and reference standard needs to be specified in the review protocol. A delay of a few days may not be a problem for chronic conditions, while for acute infectious diseases a short delay may be influential.

Signalling question 2: Did all participants receive a reference standard?

Researchers should aim to use the preferred reference standard for identifying participants with the target condition, and they should do so in all study participants. Failure to adhere to these principles could lead to bias.

Due to practical or ethical constraints, it is sometimes impossible to ascertain disease status in all participants using the preferred reference standard. If such participants remain unverified, this is known as partial verification. The biasing effect of partial verification is difficult to predict, because it depends on whether test-positive or test-negative results are not verified, whether unverified participants are omitted from the 2×2 table, whether unverified test negatives are classified as true negatives and unverified positives as true positives, and whether unverified participants can be considered random samples of index test negatives and positives (Begg 1983, Diamond 1991).

There is no correct way of handling unverified participants in an analysis; sensitivity analysis in which they are alternately considered as different combinations of test positives and test negatives may allow an assessment of the potential magnitude of any bias to be ascertained.

Partial verification is a likely threat in studies using routine care data. It may be considered good clinical practice not to perform the preferred reference standard (which can be invasive or costly) in all participants, especially in cases where the available test results indicate that the probability of the target condition being present is very low.

Random sampling of participants for verification is sometimes undertaken for reasons of efficiency, particularly in scenarios where disease prevalence is low. If participants are randomly selected to receive the reference standard (either across the whole sample or, more commonly, as random samples from index test positives and index test negatives), unbiased estimates of overall diagnostic performance of the test can be obtained if one relies on methods that compensate for the sampling plan (Zhou 1998).

Signalling question 3: Did all participants receive the same reference standard?

Some studies use an alternative reference standard in some participants, instead of a single reference standard in all. The use of different reference standards in a test accuracy study, often guided by index test results, is known as differential verification (Naaktgeboren 2013).

One clear case is when those with positive index test results receive one reference standard and those testing negative receive a different one, often a less invasive or less expensive one. For example, a study evaluating the accuracy of the D-dimer test for the diagnosis of pulmonary embolism might use CT scanning (reference standard 1) in those testing positive, but rely on clinical follow-up to decide whether or not those testing negative had a pulmonary embolism (reference standard 2). This may result in misclassifying some of the false negatives as true negatives, as some participants with a pulmonary embolism who were index test negative may be missed by clinical follow-up and so be classified as false negatives. This misclassification will over-estimate the sensitivity and specificity of the index test, compared to a study where all participants with suspected pulmonary embolism undergo CT scanning.

The studies by Lijmer et al and Rutjes et al found that the presence of differential verification was associated with an up to 2.2-fold higher diagnostic odds ratio (Lijmer 1999, Rutjes 2006).

An assessment of the risk of bias will require an understanding of the reasons for different individuals receiving different reference standards, and the nature of the differences between the reference standards.

Signalling question 4: Were all participants included in the analysis?

All participants recruited into the study should be reported and, in some form, included in the analysis. If the number of participants in the 2×2 table, used for obtaining estimates of test accuracy, differs from the number of participants included, then there is a potential for bias. Assessors should then carefully consider the intended use, context and setting of the index test, as defined in the review question, and whether the failure to include all included participants in the 2×2 table is justified.

Participants may be excluded from the 2×2 table because they had an inconclusive reference standard result. If this is a non-random subset of all participants in the study, the estimate of sensitivity generated by that study may not reflect the true proportion of test positives among those with the target condition, and the same applies to estimates of specificity.

Participants may also be excluded from the 2×2 table because of index test failures, or inconclusive index test results. The corresponding number of participants will be relevant when judging the potential usefulness of the index test in practice. Whether exclusion of the failures is justified will depend on the type of index test and the intended use, for example whether or not the test procedure can be repeated.

Whether or not a failure to include all participants in the analysis will lead to bias is a matter of judgement. A ‘No’ to this signalling question will not automatically lead to a high risk-of-bias judgement. Review authors will have to consider the proportion of participants not included in the analysis, the mechanisms for these failures and the associations, or lack thereof, with the presence or absence of the target condition. There is no universally acceptable proportion of non-included participants, although review authors may define such a proportion for their specific review in the protocol.

8.7.2 Flow and timing: additional signalling questions for comparative accuracy studies (QUADAS-C)

For comparative accuracy studies, the signalling questions should first be answered for each index test separately. There are four additional questions to consider for the flow and timing domain when dealing with comparative accuracy studies.

Additional signalling question 1: Was the risk of bias for each index test judged ‘low’ for this domain?

As discussed in the previous domains, bias in the estimates of the accuracy of an individual test may also lead to bias in the estimates of comparative accuracy of two or more index tests. Therefore, the risk of bias of each index test should be 'low' for the flow and timing domain as well.

Additional signalling question 2: Was there an appropriate interval between the index tests?

Even if the time interval between the respective index test and the reference standard is judged to be appropriate, the time interval between the index tests may not be. If the target condition is progressive, and one index test is performed at a point later in the disease progression, the accuracy of these tests may appear different only because of disease progression, even if there is no true difference in accuracy between them. The definition of appropriate interval between the index tests will be guided by the review question.

Additional signalling question 3: Was the same reference standard used for all index tests?

Ideally, in comparative accuracy studies all study participants receive the same reference standard, irrespective of the index test used. If this is not possible or feasible, then reference standards may differ between index tests. As the reference standard is a key element in the assessment of test accuracy, this will almost always affect differences in accuracy between index tests.

Additional signalling question 4: Are the proportions and reasons for missing data similar across index tests?

Differences in the number of missing data or exclusions between index tests may lead to biased estimates of comparative accuracy if missing data do not occur completely at random. For example, one index test may generate more inconclusive test results than other index test(s) being evaluated in the study. In that case, there is potential for bias if all such test results from one index test are excluded from the analysis. This question requires careful examination of the participant flow through the study and any reasons for unavailable, or inconclusive, test results as well as exclusion of participants.

Table 8.7.a summarizes the signalling questions and how these should be answered to assess the risk of bias for the flow and timing domain. Additional signalling questions are listed in Table 8.7.b.

Table 8.7.a Summary of risk-of-bias assessment for Domain 4: Flow and timing

Signalling question 1: Was there an appropriate interval between the index test and reference standard?	
Yes	If the interval between index test and reference standard was sufficiently short to avoid changes in disease status

No	If the interval is too long or too short for valid estimates of test accuracy
Unclear	If the interval between index test and reference standard was unclear
Signalling question 2: Did all participants receive a reference standard?	
Yes	If all participants received a reference standard
No	If not all participants received a reference standard
Unclear	If it is unclear whether all participants received a reference standard
Signalling question 3: Did all participants receive the same reference standard?	
Yes	If all participants received the same reference standard
No	If some participants received a different reference standard
Unclear	If it is unclear whether all participants received the same reference standard
Signalling question 4: Were all participants included in the analysis?	
Yes	If data on all study participants were included in the analysis
No	If data on all study participants were not included in the analysis
Unclear	If it is unclear whether all study participants were included in the analysis
Risk-of-bias judgement: Could participant flow have introduced bias?	
Low	If the answer to all signalling questions is ‘Yes’, then risk of bias can be considered low
High	If the answer to any of the signalling questions is ‘No’, there is a potential for bias. If one or more of the answers is ‘No’, the judgement could still be low risk of bias, but specific reasons why the risk of bias can be considered low should be provided
Unclear	If relevant information is missing for all or some of the signalling questions, and none of the answers to signalling questions is judged to put the study at high risk of bias

Table 8.7.b Comparative accuracy, additional signalling questions for the flow and timing domain (QUADAS-C)

Additional question 1: Was the risk of bias for each index test judged ‘low’ for this domain?	
Yes	If risk of bias in QUADAS-2 was judged ‘low’ for both or all index tests
No	If risk of bias in QUADAS-2 was judged to be ‘high’ or ‘unclear’ for one or more index tests
Additional question 2: Was there an appropriate interval between the index tests?	
Yes	If the interval between index tests is sufficiently short to avoid changes in disease status
No	If the interval is too long and it is likely for disease status to change in between the index tests
Unclear	If the time interval between the index tests was unclear or not reported
Additional question 3: Was the same reference standard used for all index tests?	
Yes	A single reference standard was used or, in case of differential verification, the same reference standards were used across index tests
No	The results of one index test were verified with a reference standard different to that of the other index test(s)
Unclear	It is unclear whether different reference standards were used across index tests
Additional question 4: Are the proportions and reasons for missing data similar across index tests?	
Yes	If no or only a small proportion of participant data were missing or excluded, or if the proportions and reasons for missing data were similar between the index test groups being compared
No	If more data were clearly missing from one index test group than the other(s), or the reasons for missing data differed between the index test groups
Unclear	If it is unclear whether data were missing and the extent of missing data. If data were missing, the reasons for being missing were unclear

Risk-of-bias judgement: Could the flow of participants have introduced bias in the comparison?	
Low	If the answer to all additional questions is ‘Yes’, then risk of bias can be considered low
High	If the answer to any of the signalling questions is ‘No’, there is a potential for bias. If one or more of the answers is ‘No’, the judgement could still be low risk of bias, but specific reasons why the risk of bias can be considered low should be provided
Unclear	If relevant information is missing for all or some of the signalling questions, and none of the answers to signalling questions is judged to put the study at high risk of bias

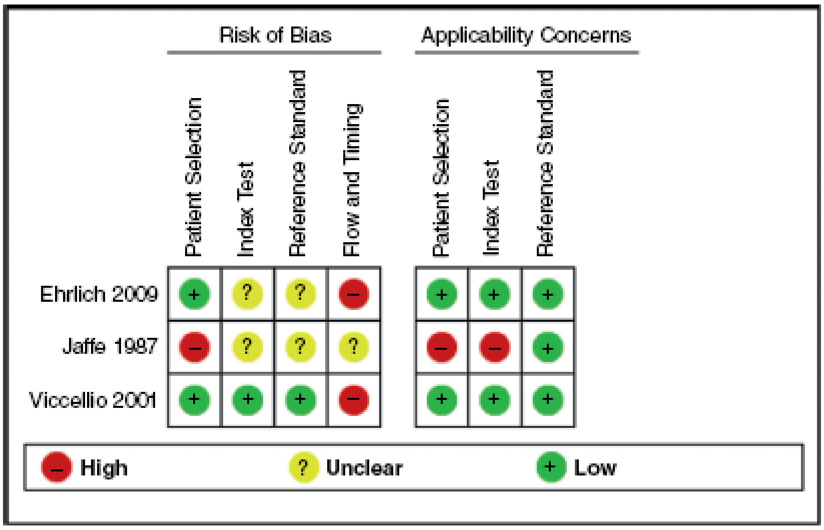
8.8 Presentation of risk-of-bias and applicability assessments

The results of the assessment of risk of bias and applicability are usually presented in systematic reviews of test accuracy using graphs or tables. The following two graphical methods provide a succinct summary of the results of the assessment – one presents a summary showing individual study results (Figure 8.8.a), while the other summarizes the results across studies (Figure 8.8.b).

The risk-of-bias and applicability concerns summary in Figure 8.8.a presents, for each included study, the ‘low’, ‘high’ and ‘unclear’ judgements for each risk-of-bias and applicability domain. This presentation works well when there are not many included studies such that the figure fits on a page and is legible.

The risk-of-bias and applicability concerns graph in Figure 8.8.b is a stacked bar chart showing the proportion of included studies for each judgement of ‘high’, ‘low’ and ‘unclear’ for each risk-of-bias and applicability domain. These two graphical methods give readers a quick overview of the risk of bias and applicability of studies included in the review.

For QUADAS-C, suggested graphical methods combining QUADAS-2 and QUADAS-C assessments are available at www.quadas.org.



	Risk of Bias				Applicability Concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Ehrlich 2009	+	?	?	–	+	+	+
Jaffe 1987	–	?	?	?	–	–	+
Viccellio 2001	+	+	+	–	+	+	+

– High ? Unclear + Low

Figure 8.8.a Risk-of-bias and applicability concerns summary table.

Source: Slaar 2017

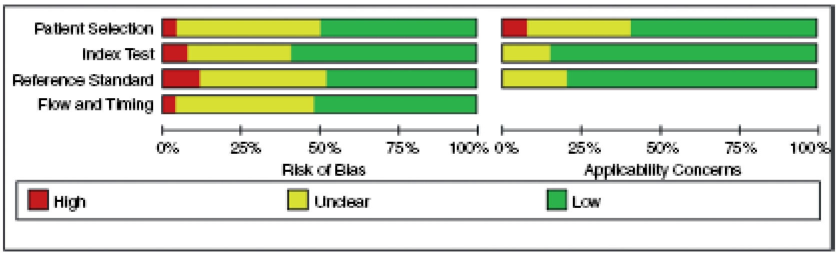


Figure 8.8.b Risk-of-bias and applicability concerns graph.

Source: Leeflang 2015

8.9 Narrative summary of risk-of-bias and applicability assesssments

In addition to presenting the results in graphs and/or tables, it is essential that a written summary of the assessment of risk of bias and applicability is included in the review. This summary should provide a narrative guide to the main potential sources of bias in the studies included in the review, and the main reasons for high risk-of-bias judgements. It may also consider the likely severity and direction of the possible biases and the nature of the applicability concerns.

In preparing a narrative summary, review authors need to incorporate:

- the frequency of ‘high’ and ‘unclear’ judgements of risk of bias and concerns regarding applicability for each domain; and

- any particular features that drive the high risk of bias and high applicability concern judgements.

For example, it is not sufficient to say that studies were at high risk of bias for participant selection; review authors should explain *why* they were at high risk of bias. This summary should facilitate a careful judgement of the degree to which the body of evidence may be compromised by bias and lack of applicability.

8.10 Chapter information

Authors: Johannes B. Reitsma (*Julius Center for Health Sciences and Primary Care, Utrecht University, The Netherlands*), Anne W. Rutjes (*Department of Medical and Surgical Sciences SMECHIMAI, University of Modena and Reggio Emilia, Italy*), Penny Whiting (*Population Health Sciences, University of Bristol, UK*), Bada Yang (*Julius Center for Health Sciences and Primary Care, Utrecht University, The Netherlands*), Mariska M. Leeflang (*Department of Epidemiology and Data Science, University of Amsterdam, The Netherlands*), Patrick M. Bossuyt (*Department of Epidemiology and Data Science, University of Amsterdam, The Netherlands*), Jonathan J. Deeks (*Institute of Applied Health Research, University of Birmingham, UK*).

Sources of support: Jonathan J. Deeks is a UK National Institute for Health Research (NIHR) Senior Investigator Emeritus. He is supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The authors declare no other sources of support for writing this chapter.

Declarations of interest: Penny Whiting is the lead author of QUADAS-2. Bada Yang is the lead author of QUADAS-C. All other authors have been involved in the development of both QUADAS-2 and QUADAS-C. Anne W. Rutjes is an Editor with Cochrane Dementia and Cognitive Improvement Group. Mariska M. Leeflang is co-convenor of the Screening and Diagnostic Test Methods Group. Mariska M. Leeflang and Jonathan J. Deeks are members of Cochrane's Diagnostic Test Accuracy Editorial Team. The authors declare no other potential conflicts of interest relevant to the topic of this chapter.

Acknowledgements: The authors thank Vasiliy Vlassov and Anne Eisinga for their contribution to previous versions of this chapter. The authors would like to thank Jacqueline Dinnes, Danielle van der Windt and Trevor McGrath for helpful peer review comments.

8.11 References

Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983; **39**: 207–215.

Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. *Lancet* 1997; **350**: 185–186.

Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, Moons KG, Reitsma JB. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS Medicine* 2013; **10**: e1001531.

Biesheuvel C, Irwig L, Bossuyt P. Observed differences in diagnostic test accuracy between patient subgroups: is it real or due to reference standard misclassification? *Clinical Chemistry* 2007; **53**: 1725–1729.

Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG. Advantages of the nested case-control design in diagnostic research. *BMC Medical Research Methodology* 2008; **8**: 48.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003a; **326**: 41–44.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, Lijmer JG. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Annals of Internal Medicine* 2003b; **138**: W1–W12.

Boyko EJ, Alderman BW, Baron AE. Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. *Journal of General Internal Medicine* 1988; **3**: 476–481.

Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, de Vet HC, Bossuyt PM. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016; **6**: e012799.

Davison BA, Harrison SA, Cotter G, Alkhouri N, Sanyal A, Edwards C, Colca JR, Iwashita J, Koch GG, Dittrich HC. Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. *Journal of Hepatology* 2020; **73**: 1322–1332.

Diamond GA. Affirmative actions: can the discriminant accuracy of a test be determined in the face of selection bias? *Medical Decision Making* 1991; **11**: 48–56.

Glasziou P, Irwig L, Deeks JJ. When should a new test become the current reference standard? *Annals of Internal Medicine* 2008; **149**: 816–822.

Gluud LL, Thorlund K, Gluud C, Woods L, Harris R, Sterne JA. Correction: reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of Internal Medicine* 2008; **149**: 219.

Higgins JPT, Altman DG. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester (UK): John Wiley & Sons; 2008: 187–241.

Ioannidis JP, Lau J. Can quality of clinical trials and meta-analyses be quantified? *Lancet* 1998; **352**: 590–591.

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled Clinical Trials* 1996; **17**: 1–12.

Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of Internal Medicine* 2001; **135**: 982–989.

Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. *Evidence-Based Medicine* 2014; **19**: 47–54.

Korevaar DA, Wang J, van Enst WA, Leeflang MM, Hooft L, Smidt N, Bossuyt PM. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology* 2015; **274**: 781–789.

Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clinical Chemistry* 2008; **54**: 729–737.

Leeflang MM, Debets-Ossenkopp YJ, Wang J, Visser CE, Scholten RJ, Hooft L, Bijlmer HA, Reitsma JB, Zhang M, Bossuyt PM, Vandenbroucke-Grauls CM. Galactomannan detection for invasive aspergillosis in immunocompromised patients. *Cochrane Database of Systematic Reviews* 2015; **12**: CD007394.

Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999; **282**: 1061–1066.

Lumbreras-Lacarra B, Ramos-Rincon JM, Hernandez-Aguado I. Methodology in diagnostic laboratory test research in clinical chemistry and clinical chemistry and laboratory medicine. *Clinical Chemistry* 2004; **50**: 530–536.

Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. Current issues and future directions. *International Journal of Technology Assessment in Health Care* 1996; **12**: 195–208.

Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine* 2002; **137**: 598–602.

Naaktgeboren CA, Bertens LCM, Smeden Mv, Groot JAHd, Moons KGM, Reitsma JB. Value of composite reference standards in diagnostic research. *BMJ* 2013; **347**: f5605.

Naylor CD. Meta-analysis and the meta-epidemiology of clinical research. *BMJ* 1997; **315**: 617–619.

Pai M, Flores LL, Pai N, Hubbard A, Riley LW, Colford JM, Jr. Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: a systematic review and meta-analysis. *Lancet Infectious Diseases* 2003; **3**: 633–643.

Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *Journal of the National Cancer Institute* 2008; **100**: 1432–1438.

Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine* 1978; **299**: 926–930.

Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995; **274**: 645–651.

Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of Clinical Epidemiology* 2009; **62**: 797–806.

Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clinical Chemistry* 2005; **51**: 1335–1341.

Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *Canadian Medical Association Journal* 2006; **174**: 469–476.

Slaar A, Fockens MM, Wang J, Maas M, Wilson DJ, Goslings JC, Schep NW, van Rijn RR. Triage tools for detecting cervical spine injury in pediatric trauma patients. *Cochrane Database of Systematic Reviews* 2017; **12**: CD011686.

Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, Bouter LM, de Vet HC. Quality of reporting of diagnostic accuracy studies. *Radiology* 2005; **235**: 347–353.

Stengel D, Bauwens K, Rademacher G, Mutze S, Ekkernkamp A. Association between compliance with methodological standards of diagnostic research and reported test accuracy: meta-analysis of focused assessment of US for trauma. *Radiology* 2005; **236**: 102–111.

Sterne JA, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Statistics in Medicine* 2002; **21**: 1513–1524.

Tullus K. Defining urinary tract infection by bacterial colony counts: a case for less than 100,000 colonies/mL as the threshold. *Pediatric Nephrology* 2019; **34**: 1651–1653.

van Rijkom HM, Verdonschot EH. Factors involved in validity measurements of diagnostic tests for approximal caries-a meta-analysis. *Caries Research* 1995; **29**: 364–370.

Verhagen AP, de Vet HC, de Bie RA, Boers M, van den Brandt PA. The art of quality assessment of RCTs included in systematic reviews. *Journal of Clinical Epidemiology* 2001; **54**: 651–654.

Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* 2003; **3**: 25.

Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *Journal of Clinical Epidemiology* 2005; **58**: 1–12.

Whiting P, Harbord R, Main C, Deeks JJ, Filippini G, Egger M, Sterne JA. Accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis: systematic review. *BMJ* 2006; **332**: 875–884.

Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011; **155**: 529–536.

Whiting PF, Rutjes AW, Westwood ME, Mallett S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology* 2013; **66**: 1093–1104.

Worster A, Carpenter C. Incorporation bias in studies of diagnostic tests: how to avoid being biased about bias. *Canadian Journal of Emergency Medicine* 2008; **10**: 174–175.

Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF, Deeks JJ, Leeflang MMG, Bossuyt PMM, Brazzelli MG, Dinnes J, Gurusamy KS, Jones HE, Lange S, Langendam MW, Macaskill P, McInnes MDF, Reitsma JB, Rutjes AWS, Sinclair A, de Vet HCW, Virgili G, Wade R, Westwood ME. QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. *Annals of Internal Medicine* 2021a; **174**: 1592–1599.

Yang B, Olsen M, Vali Y, Langendam MW, Takwoingi Y, Hyde CJ, Bossuyt PMM, Leeflang MMG. Study designs for comparative diagnostic test accuracy: a methodological review and classification scheme. *Journal of Clinical Epidemiology* 2021b; **138**: 128–138.

Zhao S, Wang S, Pan P, Xia T, Chang X, Yang X, Guo L, Meng Q, Yang F, Qian W, Xu Z, Wang Y, Wang Z, Gu L, Wang R, Jia F, Yao J, Li Z, Bai Y. Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: a systematic review and meta-analysis. *Gastroenterology* 2019; **156**: 1661–1674.

Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. *Statistical Methods in Medical Research* 1998; **7**: 337–353.