Appendix 4. Checklist for assessing the quality of diagnostic studies (QUADAS [1,2]

Author Year Article number

The checklist consists of 11 items [2]. How the different types of bias can influence the results is shown in Table 7.2 of this handbook and in the explanations/comments.

	Yes	**No**	**Unclear**
1. Was the spectrum of patients representative of those who will receive the test in practice? (representative patient spectrum)	( )	( )	( )
2. Is the reference standard likely to classify the target condition correctly?	( )	( )	( )
3. Is the interval between the reference standard and the index test sufficiently short to be reasonably certain that the target condition did not change between the two? (acceptable delay between tests)	( )	( )	( )
4. Was it the whole sample or just a random selection that was verified using a reference standard of diagnosis? (partial verification avoided)	( )	( )	( )
5. Did patients receive the same reference standard regardless of the index test result? (differential verification avoided)	( )	( )	( )
6. Was the reference standard independent of the index test (i.e. the index test was not a part of the reference standard)? (incorporation avoided)	( )	( )	( )
7. Were the reference standard results interpreted without knowledge of the index test's results?	( )	( )	( )
8. Were the index test results interpreted without knowledge of the reference standard's results?			
9. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	( )	( )	( )
10. Were uninterpretable/intermediate test results reported?	( )	( )	( )
11. Were withdrawals from the study explained?	( )	( )	( )

Explanations/comments on the individual items, and instructions for how each question in the check list should be assessed and coded[1]

1. Was the spectrum of patients representative of those who will receive the test in practice?

This question has two aspects:

- Was the right *group of patients recruited* to answer the question?
- Was the *method for recruiting patients* adequate for achieving a representative selection?

Right group of patients. It is important to have an appropriate spectrum of patients for the test under investigation since differences in demographic and clinical features between populations may lead to considerable differences between measures of diagnostic accuracy; this is known as spectrum bias. If the spectrum of tested patients is not similar to the patients for whom the test will be used in practice, the results will not be relevant. Note that a study can have good internal quality even though it is irrelevant for the question at issue. "Spectrum" refers not only to the severity of the underlying target condition, but also to demographic features and the presence of differential diagnosis and/or co-morbidity. Clinical information about the patients, such as symptoms and possible previous tests, can be essential for judging whether the composition of patients is relevant. Finally, it is important that there is a clear description of the population and that clear criteria for inclusion and exclusion of patients are reported. Information about this should be provided in reported inclusion-and exclusion criteria and/or tables concerning patient characteristics.

Theoretically, sensitivity and specificity are independent of the disease's prevalence in the population. In practice, however, both are influenced by the patient characteristics, that is, the patient spectrum to which the test is applied. This means that the sensitivity and specificity of a test applied to patients referred to a specialist clinic cannot be expected to be same as for patients in primary care. The former are a selected population, often with symptoms of disease (not yet diagnosed), while the latter are primarily less selected and have a larger proportion of patients without disease. It follows that sensitivity and specificity are also influenced by disease prevalence, which will be lower in the unselected population [3]. As a rule, sensitivity is lower and specificity higher in an unselected population [3-5].

[1] After Whiting [1] and Reitsma [2].

Method for recruiting patients

Relevance
This item is always relevant and should always be included in the quality assessment.

Specify. The reviewers should specify the relevant patient groups in advance, based on the question (inclusion criteria). Make a note of which factors may influence diagnostic accuracy, such as the clinical situation (primary care, specialist care, hospital care), the seriousness of the disease, disease prevalence and any tests done before the test in question. If a small proportion of irrelevant patients is accepted, the size of that proportion should be reported. Note also whether studies using a control group of healthy patients are accepted.

How to score this item. Studies should score "yes" for this item if you consider, on the basis of the information reported or obtained from the study's authors that the spectrum of patients included in the study was representative of those for whom the test will be used in practice. The judgment should be based on the method of recruitment as well as on the characteristics of those recruited. Studies that recruit a group of healthy controls and a group known to have the target disorder will nearly always be coded "no" on this item. The protocol of the review should specify in advance which spectrum of patients would be acceptable, taking factors such as disease prevalence and severity, age, and sex, into account. If you consider that the studied population does not come up to what you specified as acceptable, the item should be scored "no". If there is a lack of information for making this judgment, the item should be scored "unclear".

2. Is the reference standard likely to classify the target condition correctly?

The reference standard is used to determine the presence or absence of the target condition. Estimates of test performance are based on the assumption that the index test is being compared to a reference standard that is 100 per cent sensitive and specific. If there are any disagreements between the reference standard and the index test, the index test is assumed to be incorrect. The reference standard is therefore an important determinant of a test's diagnostic accuracy. Perfect reference tests are unfortunately uncommon. An imperfect reference test can result in bias concerning the index test's accuracy. As a rule, inclusion in a review should be restricted to studies that are based on one or more acceptable reference tests.

If there are serious concerns, such as the index test possibly being better than the available reference test, the usual calculations of diagnostic accuracy are no longer applicable. Under such circumstances, diagnostic accuracy cannot be calculated

without first considering whether more suitable alternative methodological methods are available.

Relevance. This item is always relevant in studies of diagnostic accuracy and should always be included in the quality assessment.

Specify. The acceptable reference standard must be defined (inclusion criteria). In some fields the reference standard is decided by consensus. A combination of reference standards is sometimes used and it may then be necessary to consider whether they are all acceptable.

How to score this item. Assessing the accuracy of a reference standard may not be straightforward. Experience of the topic area may be needed to know whether a test is an appropriate reference standard; if a combination of tests is used, the question of their appropriateness may call for careful consideration. If it is considered likely that the reference standard will classify the target condition correctly or is the best method available, this item should be scored “yes”. If this is not considered likely, the item should be scored “no”. If there is insufficient information to make a judgment, the item should be scored “unclear”.

3. Was the interval between the reference standard and the index test sufficiently short to be reasonably certain that the target condition did not change between the two?

Ideally, the results of the index test and the reference standard are collected from the same patients at the same time. If this is not possible, the delay may lead to misclassification on account of spontaneous recovery or progression to a more advanced stage of disease. The length of the interval that may cause such bias will differ between conditions. For example, a delay of just a few days is unlikely to be a problem for chronic conditions, whereas it may be important for many infectious diseases. This type of bias may occur in chronic conditions where the reference standard involves a clinical follow-up of several years.

Relevance. This item is relevant in most situations.

Specify. Decide what should be considered a “short enough” delay between the performance of the index test and the reference standard. Decide also whether it is acceptable for the interval to exceed the “short enough” interval for a certain proportion (specify the size) of the patients.

How to score this item. A judgement has to be made about what is a “short enough” interval between the index test and the reference standard, that is, the risk of wrong classification. The interval between the index test and the reference test often differs between patients. Use the longest interval between the tests to judge whether it is short enough for scoring “yes” on this item. If it is not, the score is “no”. If it is not possible assess the risk of wrong classification, the score is “unclear”.

4. Was it the whole sample or just a random selection that was verified using a reference standard?

Partial verification bias (also known as work-up bias, (primary) selection bias, or sequential ordering bias) occurs when only a ,part of the study group has the diagnosis confirmed by the reference standard. If the index test’s results influence the decision to perform the reference standard, estimates of test performance may be biased.

If patients are randomly selected to receive the reference standard, the test's overall diagnostic performance is, in theory, unchanged. In most cases, however, this selection is not random, possibly leading to biased estimates of overall diagnostic accuracy.

Relevance. Partial verification bias generally occurs only in prospective diagnostic cohort studies in which patients are tested by the index test prior to the reference standard. In situations where the reference standard is assessed before the index test, the possibility of verification bias should be assessed before scoring this item.

Specify. It may be appropriate to decide on the acceptable proportion of patients who were not verified by the reference standard.

How to score this item. If it is clear from the study that all patients, or a random selection, who received the index test went on to have their disease status verified with a reference standard, this item should be scored "yes". If some of the patients who received the index test did not have their true disease state verified and the selection of patients who received the reference standard was not random, this item should be scored "no". If this information is not reported in the study, the score should be "unclear".

5. Did patients receive the same reference standard regardless of the index test result?

Differential verification bias occurs when some of the index test results are verified by a different reference standard. This is especially a problem if these reference standards differ in their definition of the target condition. This usually occurs when patients testing positive on the index test receive a more accurate, often invasive, reference standard than those with a negative test result. Such a situation can occur when it is considered unethical to use an invasive reference test in patients with a negative index-test result. If a negative result is verified by a less accurate reference standard, measurements of test accuracy will be affected in much the same way as for partial verification, but less seriously. An extreme form of differential verification is when part of the negative test result is not verified at all. This leads to overestimation of both sensitivity and specificity.

Differential verification can also occur when different centres use different reference standards.

Empirical studies have shown that differential verification is an important source of bias [7,8]. To assess the risk of serious bias it is important to understand why different patients are verified by different reference tests and the difference in quality between reference tests. If the choice is related to the index test's results or to the probability of disease (or the condition in question), bias is a real possibility.

Relevance. Differential verification bias is a possibility in all types of diagnostic accuracy study.

Specify. As a rule, no details are necessary.

How to score this item. If it is clear that all the patients had their true disease status verified with the same reference standard, this item should be scored "yes". If a different reference standard was used for some patients, this item should be scored "no". If the study does not report this information, the score should be "unclear".

6. Was the reference standard independent of the index test (i.e. the index test was not a part of the reference standard)?

Sometimes the reference standard is decided from several components or is based on information collected over a relatively long period, for example a diagnosis of a patient released from hospital. When the index test's result is also incorporated in the basis for establishing the diagnosis (reference standard), the value of the index test will be overestimated (incorporation bias). One example is a study investigating MRI (magnetic resonance imaging) for the diagnosis of multiple sclerosis. The reference standard, the final diagnosis, was composed of all available information, including the results of MRI, cerebrospinal liquid analysis (CFC) and clinical follow-up.

Relevance. This item will only apply when a composite reference standard is used to verify disease status. In such cases it is essential that the study provides a full definition of how disease status is verified and which tests are included in the reference standard. For studies that use a single reference standard, this item will not be relevant and should be either scored "yes" or excluded removed from the quality assessment.

Specify. As a rule, no details are necessary.

How to score this item. If it is clear from the study that the index test was not a part of the reference standard, this item should be scored "yes". If the index test is considered to have been a part of the reference standard, this item should be scored "no". If this information is not reported by the study, the score should be "unclear".

7 and 8. Were the index test results interpreted without knowledge of the reference standard's results? Were the reference standard results interpreted without knowledge of the index test's results?

This item is similar to "blinding" in intervention studies. Knowledge of the results of the alternative test may influence the interpretation of the test results. This is known as review bias, and may lead to inflated measures of diagnostic accuracy. The extent to which this affects test results will be related to the degree of subjectiveness in the interpretation of the test result. The more subjective the interpretation, the more likely it is that the interpreter can be influenced by the reference standard's results in interpreting the index test and vice versa. It is therefore important to consider the topic area that is being reviewed and determine whether the interpretation of the index test or reference standard could be influenced by knowledge of the results of the other test. On some occasions, for example when laboratory tests are sent to an independent laboratory, it can be assumed that the test is interpreted independently of the reference test. However, confirmation on blinding from authors is always desirable.

Relevance. This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment. When the test results are completely objective (e.g. measurement values) or the assessment is made at an independent laboratory, the risk of review bias is small.

Specify. As a rule, no details are necessary.

How to score these items. If the study clearly states that the test results (index or reference standard) were interpreted blind to the results of the other test, these items should be scored "yes". If this does not appear to be the case, they should be scored "no". If this information is not reported by the study, the items should be scored "unclear".

9. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?

The availability of clinical data (e.g. age, presence and severity of symptoms, or other test results) during the interpretation of test results may affect estimates of test performance. One example is the interpretation of images that may be influenced by knowledge of the presence, character and localization of symptoms. If clinical data will be available when the test is interpreted in practice, they should also be available when the test is evaluated. The diagnostic value of existing clinical information before the test is performed can be difficult to separate from the added value of the index test. How this should be handled has to be based on the issue in question. Access to clinical information for the person who evaluates a test (mainly concerns radiographic images) increases sensitivity and decreases specificity [3,9].

If the index test is intended to replace other clinical tests, the results from those tests should not be available to those who interpret the index test.

Relevance. This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment.

*Specify. Make a n*ote of the clinical data that would normally be available in practice when the test is interpreted, alternatively that no information is usually available.

How to score this item. If clinical data would normally be available when the test is interpreted in practice and similar data were available when the index test was interpreted in the study, this item should be scored "yes". Similarly, if clinical data would not be available in practice and these data were not available when the index test was interpreted, this item should likewise be scored "yes". If this is not the case, this item should be scored "no". If the study does not report this information, the score should be "unclear".

10. Were uninterpretable/intermediate test results reported?

A diagnostic test can produce an uninterpretable/indeterminate result. The frequency of this varies with the nature of the test. Diagnostic accuracy studies seldom report these problems; the uninterpretable results are simply excluded from the analysis. This may lead to a biased assessment of the test's characteristics. Whether bias will arise depends on the possible correlation between uninterpretable test results and the true disease status. Whatever the cause of uninterpretable results, it is important that they are reported so that their impact on test performance can be determined.

Relevance. This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment.

Specify. As a rule, no details are necessary.

How to score this item. If it is clear that all test results, including uninterpretable/indeterminate, are reported, this item should be scored "yes". If you consider that such results occurred but have not been reported, this item should be scored "no". If it is not clear whether all study results have been reported, this item should be scored "unclear".

11. Were withdrawals from the study explained?

This occurs when patients withdraw from the study before the results of the index test and/or the reference standard are known. If patients lost to follow-up differ systematically from those who remain, for whatever reason, estimates of test performance may be biased.

Relevance. This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment.

Specify. As a rule, no details are necessary.

How to score this item. If it is clear what happened to all the patients who entered the study, for example if a flow diagram of study participants is reported, this item should be scored "yes". If it appears that some of the participants who entered the study did not complete it, i.e. did not receive both the index test and the reference standard, and these patients were not accounted for, this item should be scored "no". If it is not clear whether all patients who entered the study were accounted for, this item should be scored "unclear".

References

1. 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 Med Res Methodol 2003;3:25.

2. Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Hand-book for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration, 2009. http://srdta.cochrane.org/

3. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189-202.

4. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009;62:5-12.

5. Knottnerus JA. Diagnostic prediction rules: principles, requirements and pitfalls. Prim Care 1995;22:341-363.

6. Glasziou P, Irwig L, Deeks JJ. When should a new test become the current reference standard? Ann Intern Med 2008;149:816-22.

7. 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;282:1061-6.

8. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469-76.

9. Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical information: a systematic review. JAMA 2004; 292:1602-9.