Questions with explanations and examples

1. Was the spectrum of animals representative of the animals that will receive the test in practice?

a. What is meant by this item:
Differences in demographic, husbandry and severity of disease features between populations may produce measures of diagnostic accuracy that vary considerably. The item refers more to the generalisability or external validity of the results than to the possibility that the study may produce biased results. Reported estimates of diagnostic accuracy may have limited applicability (generalisability) if the spectrum of tested animals is not similar to the animals in which the test will be used in practice. It is therefore important that diagnostic test evaluations include an appropriate spectrum of animals for the test under investigation and also that a clear description is provided of the population of animals actually included in the study.

b. Situations in which this item does not apply:
This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool.

c. How to score this item:
Studies should score "yes" for this item if you believe, based on the information reported or obtained from the study's authors that the spectrum of animals included in the study was representative of most animals in whom the test will be used in practice in GB. The judgement should be based on both the method of recruitment and the characteristics of those recruited.

In other words, at a minimum, the study population should include a random sample of beef or dairy cattle from a random sample of herds in GB. Studies which recruit a group of healthy controls and a group known to have the target disorder or be more likely to have the target disorder will be coded as "no" on this item in nearly all circumstances. For example, a study in which the population sample of animals only included reactors to the tuberculin skin test, or animals recruited from a herd at a short interval test would be scored as no. An assessment of specificity where the prevalence of "environmental" mycobacteria is likely to be substantially different to the rest of GB would be scored as no. Studies of cattle outside GB should be scored as no.

If you think that the cattle population studied does not fit into what you specified as acceptable, the item should be scored as "no". If there is insufficient information available to make a judgement then the item should be scored as "unclear".

2. Were selection criteria clearly described?

a. What is meant by this item:
This refers to whether studies have provided a clear definition of the criteria used as inclusion and exclusion criteria for selection of animals for entry into the study.

b. Situations in which this item does not apply:
This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool.

c. How to score this item:
If relevant information regarding how animals were selected for inclusion in the study has been provided then this item should be scored as "yes".

---

At a minimum, the reference should indicate, the region from which the cattle were selected, the number of herds they were selected from and the sampling method used to select the animals. If study selection criteria are not clearly reported then this item should be scored as "no". In situations where selection criteria are partially reported and you feel that you do not have enough information to score this item as "yes", then it should be scored as "unclear".

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

a. What is meant by this item:
The reference standard is the method used to determine the presence or absence of the target condition i.e. bovine tuberculosis. To assess the diagnostic accuracy of the index test its results are compared with the results of the reference standard; subsequently indicators of diagnostic accuracy can be calculated. The reference standard is therefore an important determinant of the diagnostic accuracy of a test. Estimates of test performance are based on the assumption that the index test is being compared to a reference standard which is 100% sensitive and specific. If there are any disagreements between the reference standard and the index test then it is assumed that the index test is incorrect. Thus, from a theoretical point of view the choice of an appropriate reference standard is very important.

b. Situations in which this item does not apply:
This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool.

c. How to score this item:
If you believe that the reference standard is likely to correctly classify the target condition or is the best method available, then this item should be scored "yes". Making a judgement as to the accuracy of the reference standard may not be straightforward. If you do not think that the reference standard was likely to have correctly classified the target condition then this item should be scored as "no". If there is insufficient information to make a judgement then this should be scored as "unclear".

It is rare for a reference standard to be 100% sensitive or 100% specific so this item is likely to be scored as no on most occasions, exceptions being studies which demonstrate clearly that the animals came from a demonstrably disease-free herd (for the assessment of specificity) or that the entire animal population sample was infected with M. bovis.

4. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?

a. What is meant by this item:
Ideally the results of the index test and the reference standard are collected on the same cattle at the same time. If this is not possible and a delay occurs, misclassification due to spontaneous recovery or to progression to a more advanced stage of disease may occur. This is known as disease progression bias. The length of the time period which may cause such bias will vary between conditions. For example a delay of a few days is unlikely to be a problem for chronic conditions, however, for many infectious diseases a delay between performance of index and reference standard of only a few days may be important. This type of bias may occur in chronic conditions in which the reference standard involves clinical follow-up of several years.

b. Situations in which this item does not apply:
This item is likely to apply in most situations.

c. How to score this item:
When to score this item as "yes" is related to the target condition. For conditions that progress rapidly even a delay of several days may be important. For such conditions this item should be scored "yes" if the delay between the performance of the index and reference standard is very short, a matter of hours or days. For bovine tuberculosis, disease status with regard to detection, under conditions of natural infection, is unlikely to change within two weeks and the item should generally be scored as “yes” if the reference test (or samples obtained to perform
the test) is conducted within two weeks of the index test. However, with tuberculin skin test (index test) and a blood based antibody test (reference test) this item should only be scored as yes if the blood for the reference test was taken at the same time or shortly prior to (within 2 weeks of) the skin-test. This is because a tuberculin skin-test is an in vivo test and may influence the outcome of subsequent tests.

If you think the time period between the performance of the index test and the reference standard was sufficiently long that disease status may have changed between the performances of the two tests then this item should be scored as "no". If insufficient information is provided to make an assessment this item should be scored as "unclear".

5. Did the whole sample or a random selection of the sample, receive verification using a reference standard?

a. What is meant by this item:
Partial verification bias (also known as work-up bias, (primary) selection bias, or sequential ordering bias) occurs when not all of the study group receive confirmation of the diagnosis by the reference standard. If the results of the index test or some other factor influence the decision to perform the reference standard then biased estimates of test performance may arise, eg. if the animals reacting positively to the skin test are the only ones slaughtered and subjected to the reference standard whereas those reacting negatively to the index test are assumed to be negative to the reference standard.

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

b. Situations in which this item does not apply:
Partial verification bias will not occur when all animals that enter the study are verified against the reference standard. The bias is least likely to occur in diagnostic cohort studies in which animals are tested by the index test after being classified by the reference standard.

c. How to score this item:
If it is clear from the study that all animals, or a random selection of animals, who received the index test went on to receive verification of their disease status using a reference standard and this information was used to calculate the estimates of test performance then this item should be scored as "yes". This item should be scored as yes even if the reference standard was not the same for all animals (for example, if some cattle had macroscopic lesions and others were culture positive). If some of the patients who received the index test did not receive verification of their true disease state, and the selection of patients to receive the reference standard was not random, then this item should be scored as "no". If this information is not reported by the study then it should be scored as "unclear".

6. Did the animals receive the same reference standard regardless of the index test result?

a. What is meant by this item:
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, for example post mortem inspection and culture for the detection of infection with *M. bovis*. It may also occur when animals testing positive on the index test are likely to receive a more accurate, reference standard than those with a negative test result, take for example, a study of the sensitivity of post mortem (index test) in detecting *M. bovis* where culture for *M. bovis* (the reference standard) in cattle with macroscopic lesions is compared to culture for *M. bovis* in tissues without macroscopic lesions.

b. Situations in which this item does not apply:
Differential verification bias is possible in all types of diagnostic accuracy studies.
c. **How to score this item:**
If it is clear that animals received verification of their true disease status using the same reference standard then this item should be scored as "yes". If the standard or conduct of the verification varied then this item should be scored as "no". If this information is not reported by the study then it should be scored as "unclear".

7. **Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?**

   a. **What is meant by this item:**
   When the result of the index test is used in establishing the final diagnosis, incorporation bias may occur. This incorporation will probably increase the amount of agreement between index test results and the outcome of the reference standard, and hence overestimate the various measures of diagnostic accuracy. It is important to note that knowledge of the results of the index test alone does not automatically mean that these results are incorporated in the reference standard. For example, a study investigating visible lesions as an index test for the diagnosis of bovine tuberculosis could have a reference standard composed of identification of lesions at post-mortem and culture for *M. bovis* in tissue samples with lesions. In this case the index test forms part of the reference standard. If the reference standard had been culture of a random sample of tissues with and without lesions then the index test would not have formed part of the reference standard.

   b. **Situations in which this item does not apply:**
   This item will only apply when a composite reference standard is used to verify disease status. In such cases it is essential that a full definition of how disease status is verified and which tests form part of the reference standard are provided. For studies in which a single reference standard is used this item will not be relevant and should either be scored as yes or be removed from the quality assessment tool.

   c. **How to score this item:**
   If it is clear from the study that the index test did not form part of the reference standard then this item should be scored as "yes". If it appears that the index test formed part of the reference standard then this item should be scored as "no". If this information is not reported by the study then it should be scored as "unclear".

8. **Was the execution of the index test described in sufficient detail to permit replication of the test?**

   a. **What is meant by this item:**
   A sufficient description of the execution of index test and the reference standard is important for two reasons. Firstly, variation in measures of diagnostic accuracy can sometimes be traced back to differences in the execution of index test or reference standard. Secondly, a clear and detailed description (or citations) is needed to implement a certain test in another setting. If tests are executed in different ways then this would be expected to impact on test performance. The extent to which this would be expected to affect results would depend on the type of test being investigated.

   b. **Situations in which this item does not apply:**
   This item is likely to apply in most situations.

   c. **How to score this item:**
   If the study reports sufficient details or citations to permit replication of the index test and reference standard then these items should be scored as "yes". Studies that are from the peer reviewed scientific press should be scored as yes unless there is strong evidence that the test were performed incorrectly. In other cases these items should be scored as "no". In situations where details of test performance are partially reported and you feel that you do not have enough information to score this item as "yes", then it should be scored as "unclear".

9. **Was the execution of the reference standard described in sufficient detail to permit its replication?**
a. What is meant by this item:
A sufficient description of the execution of index test and the reference standard is important for two reasons. Firstly, variation in measures of diagnostic accuracy can sometimes be traced back to differences in the execution of index test or reference standard. Secondly, a clear and detailed description (or citations) is needed to implement a certain test in another setting. If tests are executed in different ways then this would be expected to impact on test performance. The extent to which this would be expected to affect results would depend on the type of test being investigated.

b. Situations in which this item does not apply:
This item is likely to apply in most situations.

c. How to score this item:
If the study reports sufficient details or citations to permit replication of the index test and reference standard then these items should be scored as "yes". Studies that are from the peer reviewed scientific press should be scored as yes unless there is strong evidence that the test were performed incorrectly. In other cases these items should be scored as "no". In situations where details of test performance are partially reported and you feel that you do not have enough information to score this item as "yes", then it should be scored as "unclear".

10. Were the index test results interpreted without knowledge of the results of the reference standard?

a. What is meant by this item:
This item is similar to "blinding" in intervention studies. Interpretation of the results of the index test may be influenced by knowledge of the results of the reference standard, and vice versa. This is known as review bias, and may lead to inflated measures of diagnostic accuracy. This could occur, for example, if a post-mortem is conducted as the reference test for M. bovis infection on animals where the reactor status of animals to the tuberculin skin test (the index test) is known at the time of the post-mortem. The extent to which this may affect estimates of test performance will be related to the degree of objectivity in the interpretation of the test result. The more subjective the interpretation the more likely that the interpreter can be influenced by the results of the reference standard in interpreting the index test and vice versa. It is therefore important to consider whether the interpretation of the index test or reference standard could be influenced by knowledge of the results of the other test.

b. Situations in which this item does not apply:
If, in the topic area that you are reviewing, the index test is always performed first then interpretation of the results of the index test will usually be without knowledge of the results of the reference standard. Similarly, if the reference standard is always performed first (for example, in a diagnostic case-control study) then the results of the reference standard will be interpreted without knowledge of the index test. However, if test results can be interpreted at later date, after both the index test and reference standard have been completed, then it is still important for a study to provide a description of whether the interpretation of each test was performed blind to the results of the other test. In situations where one form of review bias does not apply there are two possibilities: either score the relevant item as "yes" or remove this item from the list. If tests are entirely objective in their interpretation then test interpretation is not susceptible to review bias. In such situations review bias may not be a problem and these items can be omitted from the quality assessment tool. Another situation in which this form of bias may not apply is when tests results are interpreted in an independent laboratory. In such situations it is unlikely that the person interpreting the test results will have knowledge of the results of the other test (either index test or reference standard).

c. How to score this item:
If the study clearly states that the test results (index or reference standard) were interpreted blind to the results of the other test then these items should be scored as "yes". If this does not appear to be the case they should be scored as "no". If this information is not reported by the study then it should be scored as "unclear".
11. Were the reference standard results interpreted without knowledge of the results of the index test?

a. What is meant by this item:
This item is similar to “blinding” in intervention studies. Interpretation of the results of the index test may be influenced by knowledge of the results of the reference standard, and vice versa. This is known as review bias, and may lead to inflated measures of diagnostic accuracy. This could occur, for example, if a post-mortem is conducted as the reference test for *M. bovis* infection on animals where the reactor status of animals to the tuberculin skin test (the index test) is known at the time of the post-mortem.

The extent to which this may affect estimates of test performance will be related to the degree of objectivity in the interpretation of the test result. The more subjective the interpretation the more likely that the interpreter can be influenced by the results of the reference standard in interpreting the index test and vice versa. It is therefore important to consider whether the interpretation of the index test or reference standard could be influenced by knowledge of the results of the other test.

b. Situations in which this item does not apply:
If, in the topic area that you are reviewing, the index test is always performed first then interpretation of the results of the index test will usually be without knowledge of the results of the reference standard. Similarly, if the reference standard is always performed first (for example, in a diagnostic case-control study) then the results of the reference standard will be interpreted without knowledge of the index test. However, if test results can be interpreted at later date, after both the index test and reference standard have been completed, then it is still important for a study to provide a description of whether the interpretation of each test was performed blind to the results of the other test. In situations where one form of review bias does not apply there are two possibilities: either score the relevant item as “yes” or remove this item from the list. If tests are entirely objective in their interpretation then test interpretation is not susceptible to review bias. In such situations review bias may not be a problem and these items can be omitted from the quality assessment tool. Another situation in which this form of bias may not apply is when tests results are interpreted in an independent laboratory. In such situations it is unlikely that the person interpreting the test results will have knowledge of the results of the other test (either index test or reference standard).

c. 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 then these items should be scored as “yes”. If this does not appear to be the case they should be scored as “no”. If this information is not reported by the study then it should be scored as “unclear”.

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

a. What is meant by this item:
The availability of clinical data during interpretation of test results may affect estimates of test performance. In this context clinical data is defined broadly to include any information relating to the animal obtained by direct observation such as age, sex and symptoms. The knowledge of such factors can influence the diagnostic test result if the test involves an interpretative component. If clinical data will be available when the test is interpreted in practice then this should also be available when the test is evaluated. If however, the index test is intended to replace other clinical tests then clinical data should not be available, or should be available for all index tests. It is therefore important to determine what information will be available when test results are interpreted in practice before assessing studies for this item. This item will probably not be relevant to most assessments of diagnostic tests for bovine tuberculosis because the tests are generally performed well before the animal is likely to develop clinical symptoms. However, the item may be relevant in studies of diagnostic tests for *M. bovis* in countries without routine surveillance for bovine tuberculosis.

b. Situations in which this item does not apply:
If the interpretation of the index test is fully automated and involves no interpretation then this item may not be relevant and can be omitted from the quality assessment tool.
c. 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 interpreting the index test in the study then this item should be scored as "yes". Similarly, if clinical data would not be available in practice and these data were not available when the index test results were interpreted then this item should be scored as "yes". If this is not the case then this item should be scored as "no". If this information is not reported by the study then it should be scored as "unclear".

13. Were uninterpretable/intermediate test results reported?

a. What is meant by this item:
A diagnostic test can produce an un-interpretable/indeterminate/inconclusive result with varying frequency depending on the test. These problems are often not reported in diagnostic accuracy studies with the un-interpretable results simply removed from the analysis. This may lead to the biased assessment of the test characteristics. Whether bias will arise depends on the possible correlation between un-interpretable test results and the true disease status. If un-interpretable results occur randomly and are not related to the true disease status of the individual then, in theory, these should not have any effect on test performance. Whatever the cause of un-interpretable results it is important that these are reported so that the impact of these results on the index test performance can be determined.

b. Situations in which this item does not apply:
This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool.

c. How to score this item:
If it is clear that all test results, including un-interpretable/indeterminate/inconclusive are reported then this item should be scored as "yes". If you think that such results occurred but have not been reported, or if the un-interpretable results have been included in either the positive or negative category, then this item should be scored as "no". If it is not clear whether all study results have been reported then this item should be scored as "unclear".

14. Were withdrawals from the study explained?

a. What is meant by this item:
This occurs when animals are withdrawn from the study before the results of either or both of the index test and reference standard are known. If animals, lost to follow-up, differ systematically from those who remain, for whatever reason, then estimates of test performance may be biased.

b. Situations in which this item does not apply:
This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool.

c. How to score this item:
If it is clear what happened to all animals who entered the study, for example if a flow diagram of study animals is given showing what happened to all animals, then this item should be scored as "yes". If there were no withdrawals this item should be scored as "yes". Where drop-out is shown to have occurred at random and without knowledge of the index or reference test then the study could also be scored as a yes. If it appears that some of the animals who entered the study did not complete the study, i.e. did not receive both the index test and reference standard, and these animals were not accounted for then this item should be scored as "no". If it is not clear whether all animals that entered the study were accounted for then this item should be scored as "unclear".

15. What was the source of funding for the study?

a. What is meant by this item:
Source of funding may be associated with study results or the probability of publication (Bekelman et al (2003): Single source sponsorship may be associated with outcomes or performance of the sponsor’s product (Huss et al (2007)).

b. Situations in which this item does not apply:
This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool.

c. How to score this item:
The items should be scored as one of three options:
a) Industry (e.g. test manufacturer) b) Public or charity c) Mixed or d) Not reported.

16. Do you feel that the above answers are representative of all tests analysed within this paper?

a. What is meant by this item:
Some references will report the performance of more than one diagnostic test. It is possible that the quality of the methodology used to assess one test may be different to another in the same reference.

b. Situations in which this item does not apply:
This item is relevant to all studies of the accuracy of diagnostic tests that assess more than one test within the same reference and should always be included in the quality assessment tool.

c. How to score this item:
This items should be scored as yes if a) the performance of only one index diagnostic test is reported or the reviewers is of the view that the tests are similar in every item assessing methodological quality. If the reviewer is of the view that the studies vary in methodological quality the item should be scored as no. The item should be scored as unclear if the reviewer considers that there is insufficient information with which to compare methodological quality of diagnostic test performance.

Reference