Appendix G: Methodology checklist: the QUADAS tool for studies of diagnostic test accuracy[1]

Study identification *Including author, title, reference, year of publication*				
Guideline topic:	**Review question no:**			
Checklist completed by:				
	Circle one option for each question			
Was the spectrum of participants representative of the patients who will receive the test in practice?	Yes	No	Unclear	N/A
Were selection criteria clearly described?	Yes	No	Unclear	N/A
Was the reference standard likely to classify the target condition correctly?	Yes	No	Unclear	N/A
Was the period between performance of the reference standard and the index test short enough to be reasonably sure that the target condition did not change between the two tests?	Yes	No	Unclear	N/A
Did the whole sample or a random selection of the sample receive verification using the reference standard?	Yes	No	Unclear	N/A
Did participants receive the same reference standard regardless of the index test result?	Yes	No	Unclear	N/A
Was the reference standard independent of the index test? (that is, the index test did not form part of the reference standard)	Yes	No	Unclear	N/A
Was the execution of the index test described in sufficient detail to permit its replication?	Yes	No	Unclear	N/A
Was the execution of the reference standard described in sufficient detail to permit its replication?	Yes	No	Unclear	N/A
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes	No	Unclear	N/A
Were the reference standard results interpreted without knowledge of the results of the index test?	Yes	No	Unclear	N/A
Were the same clinical data available when the test results were interpreted as would be available when the test is used in practice?	Yes	No	Unclear	N/A
Were uninterpretable, indeterminate or intermediate test results reported?	Yes	No	Unclear	N/A
Were withdrawals from the study explained?	Yes	No	Unclear	N/A

[1] Adapted from: Whiting P, Rutjes AW, Dinnes J et al. (2004) Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technology Assessment 8: 1–234

Notes on use of Methodology checklist: studies of diagnostic test accuracy

This checklist is designed for the evaluation of studies assessing the accuracy of specific diagnostic tests. It does **not** address questions of the usefulness of the test in practice, or how the test compares with alternatives. Such questions should be assessed using the checklists for studies on interventions (see appendices D, E and F).

The questions in this checklist are aimed at establishing the validity of the study under review – that is, making sure that it has been carried out carefully, and that the conclusions represent an unbiased assessment of the accuracy and reliability of the test being evaluated. Each question covers an aspect of methodology that is thought to make a difference to the reliability of a study.

Checklist items are worded so that a 'yes' response always indicates that the study has been designed and conducted in such a way as to minimise the risk of bias for that item. An 'unclear' response to a question may arise when the answer to an item is not reported, or not reported clearly. 'N/A' should be used when a study of diagnostic test accuracy cannot give an answer of 'yes' no matter how well it has been done.

Was the spectrum of participants representative of the patients who will receive the test in practice?

What is meant by this item

Differences between populations in demographic and clinical features may produce measures of diagnostic accuracy that vary considerably; this is known as spectrum bias. Reported estimates of diagnostic test accuracy may have limited clinical applicability (generalisability) if the spectrum of participants tested is not representative of the patients on whom the test will be used in practice. The spectrum of participants takes into account not only the severity of the underlying target condition but also demographic features and the presence of differential diagnoses and/or comorbidities.

How to score this item

Studies should score 'yes' for this item if you believe, based on the information reported, that the spectrum of participants included in the study was representative of those in whom the test will be used in practice. This judgement should be based on both the method for recruitment and the characteristics of those recruited. Studies that recruited a group of healthy controls and a group known to have the target disorder will be coded as 'no' on this item in nearly all circumstances. Reviewers should pre-specify what spectrum of participants would be acceptable, taking into account factors such as disease prevalence and severity, age and sex. Clinical input may be required from the Guideline Development Group (GDG). If you think that the population studied does not fit into what you specified as acceptable, the study should be scored as 'no'. If there is insufficient information available to make a judgement, this item should be scored as 'unclear'.

Were selection criteria clearly described?

What is meant by this item

This refers to whether studies have reported criteria for entry into the study.

How to score this item

If you think that all relevant information regarding how participants were selected for inclusion in the study has been provided, then this item should be scored as 'yes'. 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'.

Was the reference standard likely to classify the target condition correctly?

What is meant by this item

The reference standard is the method used to determine the presence or absence of the target condition. Indicators of diagnostic test accuracy are calculated by comparing the results of the index test with the results of the reference standard. Estimates of test performance are based on the assumption that the index test is being compared with a reference standard that is 100% sensitive and specific. If there are any disagreements between the reference standard and the index test, it is assumed that the index test is incorrect. Thus the use of an inappropriate reference standard can bias estimation of the diagnostic accuracy of the index test.

How to score this item

Making a judgement about the accuracy of the reference standard may not be straightforward. You may need to consult a member of the GDG to determine whether a test is an appropriate reference standard. If a combination of tests is used, you may have to consider carefully whether these were appropriate.

If you believe that the reference standard is likely to classify the target condition correctly, then this item should be scored as 'yes'. If you do not think that the reference standard is likely to have classified the target condition correctly, then this item should be scored as 'no'. If there is insufficient information to make a judgement, then it should be scored as 'unclear'.

Was the period between performance of the reference standard and the index test short enough to be reasonably sure that the target condition did not change between the two tests?

What is meant by this item

Ideally, the results of the index test and the reference standard are collected on the same participants at the same time. If this is not possible and there is a delay, misclassification may occur because of either spontaneous recovery or progression of the disease. This is known as disease progression bias. The length of the period that 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 infectious diseases a delay of only a few days between performance of the index test and the reference standard may be important. This type of bias may also occur in chronic conditions in which the reference standard involves clinical follow-up of several years.

You will have to make judgements about what is considered 'short enough'. You should think about this **before** beginning your review, and define what you consider to be short enough for the specific topic area that you are reviewing. You may need clinical input to decide this.

How to score this item

When to score this item as 'yes' is related to the target condition. For conditions that progress rapidly, a delay of a even few days may be important. For such conditions this item should be scored as 'yes' if the delay between the performance of the index test and the reference standard is very short – a matter of hours or days. However, for chronic conditions, disease status is unlikely to change in a week, a month or even longer. For such conditions, longer delays between performance of the index test and reference standard may be scored as 'yes'. If you think that the period between the performance of the index test and the reference standard was sufficiently long that disease status may have changed between the performance of the two tests, then this item should be scored as 'no'. If insufficient information is provided, it should be scored as 'unclear'.

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

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 a reference standard. If the results of the index test influence the decision to perform the reference standard, then biased estimates of test performance may arise. If participants are randomly selected to receive the reference standard, the overall diagnostic performance of the test is, in theory, unchanged. However, in most cases this selection is not random, possibly leading to biased estimates of the overall diagnostic accuracy. Partial verification bias generally only occurs in diagnostic cohort studies in which participants are tested using the index test before the reference standard.

How to score this item

If it is clear from the study that all participants (or a random selection) who received the index test went on to receive verification of their disease status using a reference standard, even if this reference standard was not the same for all participants, then this item should be scored as 'yes'. If some of the participants who received the index test did not receive verification of their true disease state (or the selection was not random), then this item should be scored as 'no'. If this information is not reported, this item should be scored as 'unclear'.

Did participants receive the same reference standard regardless of the index test result?

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 a particular problem if these reference standards differ in their definition of the target condition; for example, histopathology of the appendix and natural history for the detection of appendicitis. This usually occurs when participants who test positive on the index test undergo a more accurate, often invasive, reference standard test than those with negative results on the index test. The link (correlation) between a particular (negative) test result and being verified by a less accurate reference standard can lead to biased estimates of test accuracy. Differential verification bias generally only occurs in diagnostic cohort studies in which all participants are tested using the index test before the reference standard is performed.

How to score this item

If it is clear that participants received verification of their true disease status using the same reference standard, then this item should be scored as ‘yes’. If some participants received verification using a different reference standard, then this item should be scored as ‘no’. If this information is not reported, this item should be scored as ‘unclear’.

Was the reference standard independent of the index test? (that is, the index test did not form part of the reference standard)

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 result in overestimation of the various measures of diagnostic accuracy. For example, a study investigating magnetic resonance imaging (MRI) for the diagnosis of multiple sclerosis could have a reference standard composed of clinical follow-up, cerebrospinal fluid analysis and MRI. In this case, the index test forms part of the reference standard. It is important to note that knowledge of the results of the index test does not automatically mean that these results are incorporated in the reference standard. 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 is provided.

How to score this item

For studies in which a single reference standard is used, this item will not be relevant and should be scored as ‘N/A’. If it is clear 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, this item should be scored as ‘unclear’.

Was the execution of the index test described in sufficient detail to permit its replication?
Was the execution of the reference standard described in sufficient detail to permit its replication?

What is meant by these items

A sufficiently detailed description of the execution of the 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 tests and reference standards. Secondly, a clear and detailed description (or references) is needed to implement a certain test in another setting. If tests are executed in different ways then this would be expected to have an impact on test performance. The extent to which this would be expected to affect results depends on the type of test being investigated.

How to score these items

If the study reports sufficient details to permit replication of the index test and the reference standard, then these items should be scored as 'yes'. In other cases these items should be scored as 'no'. In situations where details of test performance are partially reported and you consider that you do not have enough information to score these items as 'yes', then they should be scored as 'unclear'.

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

What is meant by these items

This issue is similar to the blinding of the people who assess outcomes 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 test accuracy. The extent to which this can affect test results will be related to the degree of subjectivity 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 index test in interpreting the results of the reference standard, and vice versa. It is therefore important to consider the topic area that you are reviewing and to determine whether interpretation of the results of the index test or the reference standard could be influenced by knowledge of the results of the other test.

How to score these items

If the study clearly states that the test results (index test 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, then they should be scored as 'no'. If this information is not reported, these items should be scored as 'unclear'. 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 done without knowledge of the results of the reference standard. Similarly, if the reference standard is always performed first, then the results will be interpreted without knowledge of the results of the index test. In situations where one form of review bias does not apply, the item should be scored as 'N/A'. If interpretation of test results is entirely objective, then test interpretation is not susceptible to review bias and the item should be scored as 'N/A'. Another situation in which this form of bias may not apply is when test 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).

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

What is meant by this item

The availability of information on clinical data during the interpretation of test results may affect estimates of test performance. In this context, clinical data are defined broadly to include any information relating to the participant that is 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 these should also be available when the test is evaluated. However, if the index test is intended to replace other clinical tests, then clinical data should not be available. Thus, before assessing studies for this item it is important to determine what information will be available when test results are interpreted in practice. You should consult the GDG to identify this information.

How to score this item

If clinical data would normally be available when the test results are interpreted in practice and similar data were available when interpreting the index test results 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, this item should be scored as 'unclear'. If interpretation of the index test is fully automated, this item may not be relevant and can be scored 'N/A'.

Were uninterpretable, indeterminate or intermediate test results reported?

What is meant by this item

A diagnostic test can produce an uninterpretable, indeterminate or intermediate result with varying frequency, depending on the test. These problems are often not reported in studies on diagnostic test accuracy, the uninterpretable results simply being 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 uninterpretable test results and

the true disease status. If uninterpretable 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. It is important that uninterpretable results are reported so that the impact on test performance can be considered; however, poor quality of reporting means that this is not always the case.

How to score this item

If it is clear that all test results, including uninterpretable, indeterminate or intermediate results, are reported, then this item should be scored as 'yes'. If the authors do not report any uninterpretable, indeterminate or intermediate results, and if the results are reported for all participants who were described as having been entered into the study, then this item should also be scored as 'yes'. If you think that such results occurred but have not been reported, 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'.

Were withdrawals from the study explained?

What is meant by this item

This occurs when participants withdraw from the study before the results of both the index test and the reference standard are known. If participants lost to follow-up differ systematically from those who remain, for whatever reason, then estimates of test performance may be biased. Poor quality of reporting of withdrawals may make the impact on estimates of test performance difficult to determine.

How to score this item

If it is clear what happened to all participants who entered the study, for example if a flow diagram of study participants is reported, then this item should be scored as 'yes'. If the authors do not report any withdrawals and if results are available for all participants who were reported to have been entered into the study, then this item should also be scored as 'yes'. If it appears that some of the participants who entered the study did not complete the study (that is, did not receive both the index test and the reference standard), and these participants were not accounted for, then this item should be scored as 'no'. If it is not clear whether all participants who entered the study were accounted for, then this item should be scored as 'unclear'.

BMC Medical Research Methodology

BioMed Central

Research article

Open Access

The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews

Penny Whiting*[1], Anne WS Rutjes[2], Johannes B Reitsma[2], Patrick MM Bossuyt[2] and Jos Kleijnen[1]

Address: [1]Centre for Reviews and Dissemination, University of York, England, UK and [2]Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, The Netherlands

Email: Penny Whiting* - pfw2@york.ac.uk; Anne WS Rutjes - a.rutjes@amc.uva.nl; Johannes B Reitsma - j.reitsma@amc.uva.nl; Patrick MM Bossuyt - p.m.bossuyt@amc.uva.nl; Jos Kleijnen - jk13@york.ac.uk

* Corresponding author

Published: 10 November 2003

BMC Medical Research Methodology 2003, **3**:25

Received: 14 July 2003
Accepted: 10 November 2003

This article is available from: http://www.biomedcentral.com/1471-2288/3/25

Abstract

Background: In the era of evidence based medicine, with systematic reviews as its cornerstone, adequate quality assessment tools should be available. There is currently a lack of a systematically developed and evaluated tool for the assessment of diagnostic accuracy studies. The aim of this project was to combine empirical evidence and expert opinion in a formal consensus method to develop a tool to be used in systematic reviews to assess the quality of primary studies of diagnostic accuracy.

Methods: We conducted a Delphi procedure to develop the quality assessment tool by refining an initial list of items. Members of the Delphi panel were experts in the area of diagnostic research. The results of three previously conducted reviews of the diagnostic literature were used to generate a list of potential items for inclusion in the tool and to provide an evidence base upon which to develop the tool.

Results: A total of nine experts in the field of diagnostics took part in the Delphi procedure. The Delphi procedure consisted of four rounds, after which agreement was reached on the items to be included in the tool which we have called QUADAS. The initial list of 28 items was reduced to fourteen items in the final tool. Items included covered patient spectrum, reference standard, disease progression bias, verification bias, review bias, clinical review bias, incorporation bias, test execution, study withdrawals, and indeterminate results. The QUADAS tool is presented together with guidelines for scoring each of the items included in the tool.

Conclusions: This project has produced an evidence based quality assessment tool to be used in systematic reviews of diagnostic accuracy studies. Further work to determine the usability and validity of the tool continues.

Background

Systematic reviews aim to identify and evaluate all available research evidence relating to a particular objective. [1] Quality assessment is an integral part of any systematic

review. If the results of individual studies are biased and these are synthesised without any consideration of quality then the results of the review will also be biased. It is therefore essential that the quality of individual studies included in a systematic review is assessed in terms of potential for bias, lack of applicability, and, inevitably to a certain extent, the quality of reporting. A formal assessment of the quality of primary studies included in a review allows investigation of the effect of different biases and sources of variation on study results.

Quality assessment is as important in systematic reviews of diagnostic accuracy studies as it is in any other review. [2] However, diagnostic accuracy studies have several unique features in terms of design which differ from standard intervention evaluations. Their aim is to determine how good a particular test is at detecting the target condition and they usually have the following basic structure. A series of patients receive the test (or tests) of interest, known as the "index test(s)" and also a reference standard. The results of the index test(s) are then compared to the results of the reference standard. The reference standard should be the best available method to determine whether or not the patient has the condition of interest. It may be a single test, clinical follow-up, or a combination of tests. Both the terms "test" and "condition" are interpreted in a broad sense. The term "test" is used to refer to any procedure used to gather information on the health status of an individual. This can include laboratory tests, surgical exploration, clinical examination, imaging tests, questionnaires, and pathology. Similarly "condition" can be used to define any health status including the presence of disease (e.g. influenza, alcoholism, depression, cancer), pregnancy, or different stages of disease (e.g. an exacerbation of multiple sclerosis).

Diagnostic accuracy studies allow the calculation of various statistics that provide an indication of "test performance" – how good the index test is at detecting the target condition. These statistics include sensitivity, specificity, positive and negative predictive values, positive and negative likelihood ratios, diagnostics odds ratios and receiver operating characteristic (ROC) curves.

Unique design features mean that the criteria needed to assess the quality of diagnostic test evaluations differ from those needed to assess evaluations of therapeutic interventions. [2] It is also important to use a standardised approach to quality assessment. This should avoid the choice of a quality assessment tool that is biased by preconceived ideas. [1] Although several checklists for the assessment of the quality of studies of diagnostic accuracy exist, none of these have been systematically developed or evaluated, and they differ in terms of the items that they assess. [3]

The aim of this project was to use a formal consensus method to develop and evaluate an evidence based quality assessment tool, to be used for the quality assessment of diagnostic accuracy studies included in systematic reviews.

Methods

We followed the approach suggested by Streiner and Norman to develop the quality assessment tool [4]. Jadad et al [5] also adopted this approach to establish a scale for assessing the quality of randomised controlled studies. This procedure involves the following stages: (1) preliminary conceptual decisions; (2) item generation; (3) assessment of face validity; (4) field trials to assess consistency and construct validity; and lastly (5) the generation of the refined instrument.

Preliminary conceptual decisions

We decided that the quality assessment tool was required to:

1. be used in systematic reviews of diagnostic accuracy

2. assess the methodological quality of a diagnostic study in generic terms (relevant to all diagnostic studies)

3. allow consistent and reliable assessment of quality by raters with different backgrounds

4. be relatively short and simple to complete

'Quality' was defined to include both the internal and external validity of a study; the degree to which estimates of diagnostic accuracy have not been biased, and the degree to which the results of a study can be applied to patients in practice.

We conducted a systematic review of existing systematic reviews of diagnostic accuracy studies to investigate how quality was incorporated into these reviews. The results of this review are reported elsewhere. [9,10] Based on these results, we decided that the quality assessment tool needed to have the potential to be used:

- as criteria for including/excluding studies in a review or in primary analyses

- to conduct sensitivity/subgroup analysis stratified according to quality

- as individual items in meta-regression analyses

- to make recommendations for future research

- to produce a narrative discussion of quality

• to produce a tabular summary of the results of the quality assessment

The implication for the development of the tool was that it needed to be able to distinguish between high and low quality studies. Component analysis, where the effect of each individual quality item on estimates of test performance is assessed, was adopted as the best approach to incorporate quality into systematic reviews of diagnostic studies. [6,7] We decided not to use QUADAS to produce an overall quality score due to the problems associated with their use. [8] The quality tool was developed taking these aspects into consideration.

Item generation

We produced an initial list of possible items for inclusion in the quality assessment tool incorporating the results of two previously conducted systematic reviews. Full details of these are reported elsewhere. [3,10,11] The first review examined the methodological literature on diagnostic test assessment to identify empirical evidence for potential sources of bias and variation. [11] The results from this review were summarised according to the number of studies providing empirical, theoretical or no evidence of bias. The second looked at existing quality assessment tools to identify all possible relevant items and investigated on what evidence those items are based. [3] The results from this review were summarised according to the proportion of tools that included each item, this formed the basis for the initial list. We phrased each proposed item for the checklist as a question.

Assessment of face validity

The main component of the development of the tool was the assessment of face validity. We chose to use a Delphi procedure for this component. Delphi procedures aim to obtain the most reliable consensus amongst a group of experts by a series of questionnaires interspersed with controlled feedback. [12,13] We felt that the Delphi procedure was the optimum method to obtain consensus on the items to be included in the tool as well as the phrasing and scoring of items. This method allowed us to capture the views of a number of experts in the field and to reach a consensus on the final selection of items for the tool. As each round of the procedure is completed independently, the views of each expert panel member can be captured without others influencing their choices. However, at the same time, consensus can be reached by the process of anonymously feeding back the responses of each panel member in a controlled manner in subsequent rounds.

As the area of diagnostic accuracy studies is a specialised area we decided to include a small number of experts in the field on the panel, rather than to include a larger number of participants who may have had a more limited knowledge of the area. Eleven experts were contacted and asked to become panel members.

The Delphi procedure

General features

Each round of the Delphi procedure included a report of the results from the previous round to provide a summary of the responses of all panel members. We also provided details on how we reached decisions regarding which items to include/exclude from the tool based on the results of the previous round. We reported all other decisions made, for example how to handle missing responses and rephrasing of items, together with the justification for the decisions. These decisions were made by the authors who were not panel members (PW, AR, JR, JK (the 'steering group')) and we asked panel members whether they supported the decisions made. When making a decision regarding whether to include an item in the quality assessment tool, we asked panel members to consider the results from the previous round, the comments from the previous round (where applicable), *and* the evidence provided for each item.

Delphi Round 1

We sent the initial list of possible items for inclusion in the quality assessment tool, divided into four categories (Table 1 [see Additional file 1]), to all panel members. The aim was to collect information on each member of the group's opinion regarding the importance of each item. To help panel members in their decision-making, we summarised the evidence from the reviews for each item. The aims of the quality assessment tool and its desired features were presented. We asked members of the panel to rate each item for inclusion in the quality assessment tool according to a five point Likert Scale (strongly agree, moderately agree, neutral, moderately disagree, strongly disagree). We also gave them the opportunity to comment on any of the items included in the tool, to suggest possible rephrasing of questions and to highlight any items that may have been missed off the initial list of items.

Delphi Round 2

We used the results of round 1 to select items for which there were high levels of agreement for inclusion/exclusion from the final quality assessment tool. Categories/items rated as "strongly agree" by at least 75 % of the panel members who replied in this round were selected for inclusion in the tool. Categories/items that were not rated as "strongly agree" by at least one panel member were excluded. Items selected for inclusion or exclusion from the final quality assessment tool were not rated as part of round 2.

For the round 2 questionnaire, rather than rating each item on the 5-point Likert scale, we asked panel members

to indicate whether they thought that a category or item should be included or excluded from the quality assessment tool. In addition, we asked panel members to answer yes or no to the following questions:

1. Would you like to see a number of "key items" highlighted in the quality assessment tool?

2. Do you endorse the Delphi procedure so far? If no, please give details of the aspects of the procedure which you do not support and list any suggestions you have for how the procedure could be improved.

3. As part of the third round, instructions on how to complete the quality assessment will be provided to you. As we do not want to ask you to invest too much time, the instructions will be drawn up by the steering group. In the third round you will only be asked if you support the instructions and if not, what you would like to change. Do you agree with this procedure?

We described the methods proposed to validate the tool and asked panel members to indicate whether or not they agreed with these methods, and also to suggest any additional validation methods.

Delphi Round 3

We used the results of round 2 to further select items for inclusion/exclusion in the quality assessment tool. All categories rated as include by more than 80% of the panel members were selected for inclusion in the tool. Items scored "include" by 75% of the panel members were re-rated as part of round 3. All other items were removed from the tool and comments regarding rephrasing were incorporated while revising the tool.

We presented all items selected for inclusion in the tool at this stage and asked panel members to indicate if they agreed with the proposed phrasing of the items, and if not, to suggest alternative phrasings. As for the round 2 questionnaire, we asked panel members to indicate whether they thought that each item to be re-rated should be included or excluded from the quality assessment tool.

We proposed a scoring system and asked panel members to indicate whether they agreed with this system. The system proposed was straightforward: all items would be scored as "yes", "no" or "unclear". We presented further details of the proposed validation methods and again asked panel members to indicate whether they agreed with these methods. The aims of the quality assessment tool were highlighted and we asked panel members whether taking these into consideration they endorsed the Delphi procedure. We also asked members whether they used the evidence provided from the reviews and the feedback from previous rounds in their decisions of which items to select for inclusion in the tool. If they did not use this information we asked them to explain why not.

Lastly, we asked panel members if they would like to see the development of topic and design specific items in addition to the generic section of the tool. If they answered yes to these questions we asked them whether they would like to see the development of these items through a further Delphi procedure, and if so, if they would like to be a member of the panel for this procedure. We decided to name the tool the "QUADAS" (**Qu**ality **A**ssessment of **D**iagnostic **A**ccuracy **S**tudies) tool. We produced a background document to accompany QUADAS for the items selected for inclusion in the tool up to this point and asked panel members to comment on this.

Delphi Round 4

We used the results of round 3 to select the final items for inclusion/exclusion in the quality assessment tool. All items rated as 'include' by at least 75% of the panel members were selected for inclusion in the tool. All other items were removed from the tool. We considered comments regarding rephrasing of items and rephrased items taking these into consideration. The final version of the background document to accompany QUADAS was presented.

Field trials to assess consistency and construct validity

We are currently in the process of evaluating the consistency, validity and usability of QUADAS. The validation process will include piloting the tool on a small sample of published studies, focussing on the assessment of the consistency and reliability of the tool. The tool will also be piloted in a number of diagnostic reviews. Regression analysis will be used to investigate associations between study characteristics and estimates of diagnostic accuracy in primary studies.

Generation of the refined instrument

If necessary, we will use the results of the evaluations outlined above to adapt QUADAS. The steps involved in the development of QUADAS are illustrated in Figure 1.

Results

Item generation

The systematic reviews produced a list of 28 possible items for inclusion in the quality assessment tool. These are shown in Table 1 [see Additional file 1] together with the results of the systematic reviews on sources of bias and variation, and existing quality assessment tools. The evidence from the review on sources of bias and variation was summarised as the number of studies reporting empirical evidence (E), theoretical evidence (T) or absence (A) of bias or variability. The number of studies providing each type of evidence of bias or variability is

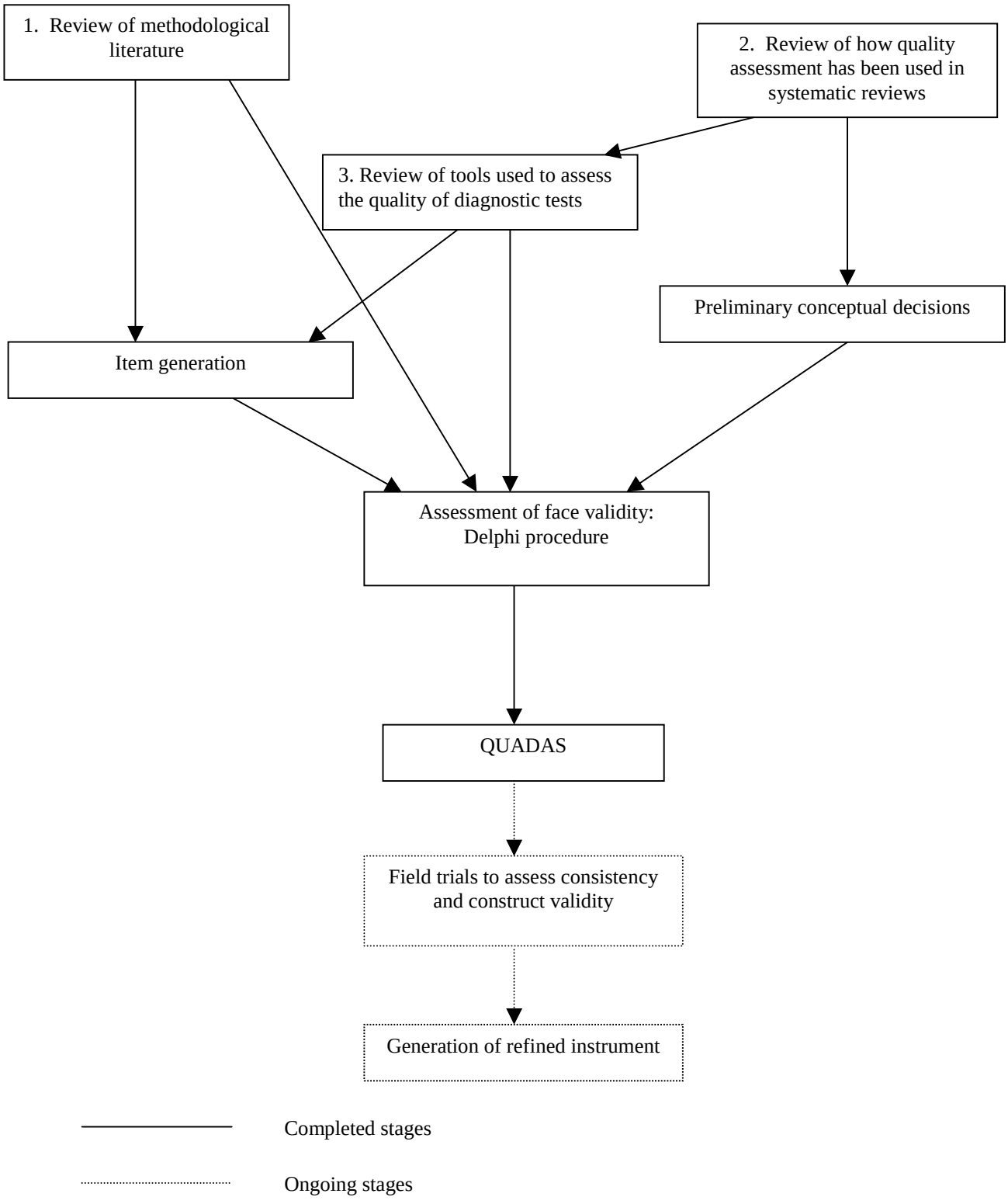

Figure 1
Flow chart of the tool development process.

shown in columns 2–4 of Table 1 [see Additional file 1]. The results from the review of existing quality assessment tools was summarised as the proportion of tools covering each item. The proportions were grouped into four categories: I (75–100%), II (50–74%), III (25–49%) and IV (0–24%) and are shown in the final column of Table 1 [see Additional file 1]. For some items evidence from the reviews was only available in combination with other items rather than for each item individually, e.g. setting and disease prevalence and severity. For these items the evidence for the combination is provided. Table 1 [see Additional file 1] also shows to which item on the QUADAS tool each item in this table refers. Evidence from the first review was not available for a number of items; these items were classed as "na".

Assessment of face validity: The Delphi procedure

Nine of the eleven people invited to take part in the Delphi procedure agreed to do so. The names of the panel members are listed at the end of this paper.

Delphi Round 1

Eight of the nine people who agreed to take part in the procedure returned completed questionnaires. The ninth panel member did not have time to take part in this round. Following the results of this round, six items were selected for inclusion, one item was removed from the tool, and the remaining items were put forward to be re-rated as part of round 2. Items selected for inclusion were:

1. Appropriate selection of patient spectrum

2. Appropriate reference standard

3. Absence of partial verification bias

4. Absence of review bias (both test and diagnostic)

5. Clinical review bias

6. Reporting of uninterpretable/indeterminate/intermediate results

The item removed from the tool was:

1. Test utility

Panel members made a number of suggestions regarding rephrasing of items. We considered these and made changes where appropriate. Based on some of the comments received we added an additional item to the category "Spectrum composition". This item was "What was the study design?". This item was rated for inclusion in the tool as part of round 2.

Delphi Round 2

Of the nine people invited to take part in round 2, eight returned completed questionnaires. Based on the results of this round, a further four items were selected for inclusion in the tool:

1. Absence of disease progression bias

2. Absence of differential verification bias

3. Absence of incorporation bias

4. Reporting of study withdrawals.

Panel members did not achieve consensus for a further five items, these were re-rated as part of round 3:

1. Reporting of selection criteria

2. Reporting of disease severity

3. Description of index test execution

4. Description of reference standard execution

5. Independent derivation of cut-off points

All other items, including the new item added based on feedback from round 1, were excluded from the process. Based on the comments from round 2, we proposed the following additional items which were included in the round 3 questionnaire:

1. Are there other aspects of the design of this study which cause concern about whether or not it will correctly estimate test accuracy?

2. Are there other aspects of the conduct of this study which cause concern about whether or not it will correctly estimate test accuracy?

3. Are there special issues concerning patient selection which might invalidate test results?

4. Are there special issues concerning the conduct of test which might invalidate test results?

Since none of the panel members were in favour of highlighting a number of key items in the quality assessment tool, this approach was not followed. At this stage, five of the panel members reported that they endorsed the Delphi procedure so far, one did not and two were unclear. The member who did not endorse the Delphi procedure stated that "I fundamentally believe that it is not possible to develop a reliable discriminatory diagnostic assessment

tool that will apply to all, or even the majority of diagnostic test studies." One of the comments from a panel member who was "unclear" also related to the problem of producing a quality assessment tool that applies to all diagnostic accuracy studies. The other related to the process used to derive the initial list of items and the problems of suggesting additional items. All panel members agreed to let the steering group produce the background document to accompany the tool. The feedback suggested that there was some confusion regarding the proposed validation methods. These were clarified and re-rated as part of round 3.

Delphi Round 3

All nine panel members invited to take part in round 3 returned completed questionnaires. Agreement was reached on items to be included in the tool following the results of this round.

Three of the five items re-rated as part of this round were selected for inclusion. These were:

1. Reporting of selection criteria

2. Description of index test execution

3. Description of reference standard execution

The other two items and the additional items rated as part of this round were not included in the tool.

The panel members agreed with the scoring system proposed by the steering group. Each of the proposed validation steps was approved by at least 7/9 of the panel members. These methods will therefore be used to validate the tool. Five of the panel members indicated that they would like to see the development of design and topic specific criteria. Of these four stated that they would like to see this done via a Delphi procedure. The development of these elements will take place after the generic section of the tool has been evaluated.

At this stage, all but one of the panel members stated that they endorsed the Delphi procedure. This member remained unclear as to whether he/she endorsed the procedure and stated that "all my reservations still apply". These reservations related to earlier comments regarding the problems of developing a quality assessment tool which can be applied to all studies of diagnostic accuracy. Seven of the panel members reported using the evidence provided from the systematic reviews to help in their decisions of which items to include in QUADAS. Of the two that did not use the evidence one stated that (s)he was too busy, the other stated that there was no new information in the evidence. Seven of the panel members reported using the feedback from earlier rounds of the Delphi procedure. Of the two that did not, one stated that he/she was "not seeking conformity with other respondents" the other did not explain why he or she did not use the feedback. The two panel members that did not use the feedback were different from the two that did not use the evidence provided by the reviews. These responses suggest that the evidence provided by the review did contribute towards the production of QUADAS.

Delphi Round 4

The fourth and final round did not include a questionnaire, although panel members were given the opportunity to feedback any additional comments that they had. Only one panel member provided further feedback. This related mainly to the broadness of the first item included in the tool, and the fact that several items relate to the reporting of the study rather than directly to the quality of the study.

The QUADAS tool

The tool is structured as a list of 14 questions which should each be answered "yes", "no", or "unclear". The tool is presented in Table 1. A more detailed description of each item together with a guide on how to score each item is provided below.

Users' guide to QUADAS

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

a. What is meant by this item

Differences in demographic and clinical features between populations may produce measures of diagnostic accuracy that vary considerably, this is known as spectrum bias. It refers more to the generalisability of results than to the possibility that the study may produce biased results. Reported estimates of diagnostic accuracy may have limited clinical applicability (generalisability) if the spectrum of tested patients is not similar to the patients in whom the test will be used in practice. The spectrum of patients refers not only to the severity of the underlying target condition, but also to demographic features and to the presence of differential diagnosis and/or co-morbidity. It is therefore important that diagnostic test evaluations include an appropriate spectrum of patients for the test under investigation and also that a clear description is provided of the population 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.

Table 2: The QUADAS tool

Item		Yes	No	Unclear
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	( )	( )	( )
2.	Were selection criteria clearly described?	( )	( )	( )
3.	Is the reference standard likely to correctly classify the target condition?	( )	( )	( )
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?	( )	( )	( )
5.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	( )	( )	( )
6.	Did patients receive the same reference standard regardless of the index test result?	( )	( )	( )
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	( )	( )	( )
8.	Was the execution of the index test described in sufficient detail to permit replication of the test?	( )	( )	( )
9.	Was the execution of the reference standard described in sufficient detail to permit its replication?	( )	( )	( )
10.	Were the index test results interpreted without knowledge of the results of the reference standard?	( )	( )	( )
11.	Were the reference standard results interpreted without knowledge of the results of the index test?	( )	( )	( )
12.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	( )	( )	( )
13.	Were uninterpretable/ intermediate test results reported?	( )	( )	( )
14.	Were withdrawals from the study explained?	( )	( )	( )

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 patients included in the study was representative of those in whom the test will be used in practice. The judgement should be based on both the method of recruitment and the characteristics of those recruited. Studies which recruit a group of healthy controls and a group known to have the target disorder will be coded as "no" on this item in nearly all circumstances. Reviewers should pre-specify in the protocol of the review what spectrum of patients would be acceptable taking factors such as disease prevalence and severity, age, and sex, into account. If you think that the 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 it 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 in- and exclusion criteria 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 you think that all relevant information regarding how participants were selected for inclusion in the study has been provided then this item should be scored as "yes". 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. 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. You may need experience of the topic area to know whether a test is an appropriate reference standard, or if a combination of tests are used you may have to consider carefully whether these were appropriate. 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".

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 patients 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. However, for chronic conditions disease status is unlikely to change in a week, or a month, or even longer. In such conditions longer delays between performance of the index and reference standard may be scored as "yes". You will have to make judgements regarding what is considered "short enough". You should think about this before starting work on a review, and define what you consider to be "short enough" for the specific topic area that you are reviewing. 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 performance of the two tests then this item should be scored as "no". If insufficient information is provided this 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 influence the decision to perform the reference standard then biased estimates of test performance may arise. If patients are randomly selected to receive the reference standard the overall diagnostic performance of the test is, in theory, unchanged. 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 generally only occurs in 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, you should firstly decide whether there is a possibility that verification bias could occur, and if not how to score this item. This may depend on how quality will be incorporated in the review. There are two options: either to score this item as 'yes', or to remove it from the quality assessment tool.

c. How to score this item

If it is clear from the study that all patients, or a random selection of patients, who received the index test went on to receive verification of their disease status using a reference standard 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 patients. 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 patients 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 stand-

ard. This is especially a problem if these reference standards differ in their definition of the target condition, for example histopathology of the appendix and natural history for the detection of appendicitis. 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. The link (correlation) between a particular (negative) test result and being verified by a less accurate reference standard will affect measures of test accuracy in a similar way as for partial verification, but less seriously.

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 patients received verification of their true disease status using the same reference standard then this item should be scored as "yes". If some patients received verification using a different reference standard 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 MRI for the diagnosis of multiple sclerosis could have a reference standard composed of clinical follow-up, CSF analysis and MRI. In this case the index test forms part of the reference standard. If the same study used a reference standard of clinical follow-up and the results of the MRI were known when the clinical diagnosis was made but were not specifically included as part of the reference then the index test does not form 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?

9. Was the execution of the reference standard described in sufficient detail to permit its replication?

a. What is meant by these items

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 these items do not apply

These items are likely to apply in most situations.

c. How to score these items

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". 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?

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

a. What is meant by these items

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. The extent to which this may affect 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 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 the topic area that you are reviewing and to determine 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 these items do 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 patient 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.

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 uninterpretable/indeterminate/intermediate result with varying frequency depending on the test. These problems are often not reported in diagnostic accuracy studies with the uninterpretable 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 uninterpretable test results and the true disease status. If uninterpretable 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 uninterpretable results it is important that these are reported so that the impact of these results on 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 uninterpretable/ indeterminate/intermediate are reported then this item should be scored as "yes". If you think that such results occurred but have not been reported 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 patients withdraw from the study before the results of either or both of the index test and reference standard are known. If patients 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 patients who entered the study, for example if a flow diagram of study participants is reported, then this item should be scored as "yes". If it appears that some of the participants who entered the study did not complete the study, i.e. did not receive both the index test and reference standard, and these patients were not accounted for then this item should be scored as "no". If it is not clear whether all patients who entered the study were accounted for then this item should be scored as "unclear".

Discussion

This project has produced an evidence based tool for the quality assessment of studies of diagnostic accuracy. The tool is now available to all reviewers involved in systematic reviews of studies of diagnostic accuracy. The final tool consists of a set of 14 items, phrased as questions, each of which should be scored as yes, no or unclear. The tool is simple and quick to complete and does not incorporate a quality score. There are a number of reasons for not incorporating a quality score into QUADAS. Quality scores are only necessary if the reviewer wants to use an overall indicator of quality to weight the meta-analysis, or as a continuous variable in a meta-regression. Since quality scores are very rarely used in these ways, we see no need to introduce such a score. Choices on how to weight and calculate quality scores are generally fairly arbitrary thus it would be impossible to produce an objective quality score. Furthermore, quality scores ignore the fact that the importance of individual items and the direction of potential biases associated with these items may vary according to the context in which they are applied. [6,7] The application of quality scores, with no consideration of the individual quality items, may therefore dilute or entirely miss potential associations. [8]

Experts in the area used evidence provided by systematic reviews of the literature on diagnostic accuracy studies to produce the quality assessment tool. This is the first tool that has been systematically developed in the field of diagnostic accuracy studies. A further strength of this tool is that it will be subjected to a thorough evaluation. Any problems with the tool highlighted by this evaluation will be addressed with the aim of improving the tool. The tool is currently being piloted in 15 reviews of diagnostic accuracy studies covering a wide range of topics including the diagnosis of multiple sclerosis, the diagnosis of urinary tract infection in children under 5, diagnosing urinary incontinence and myocardial perfusion scintigraphy. We are collecting feedback on reviewers' experience of the use of QUADAS through a structured questionnaire. Anyone interested in helping pilot the tool can contact the authors for a questionnaire. Other proposed work to evaluate the tool includes an assessment of the consistency and reliability of the tool.

There are a number of limitations to this project, and the QUADAS tool. The main problem relates to the development of a single tool which can be applied to all diagnostic accuracy studies. The objective of this project was not to produce a tool to cover everything, but to produce a quality assessment tool that can be used to assess the quality of primary studies included in systematic reviews. We appreciate that different aspects of quality will be applicable to different topic areas and for different study designs. However, QUADAS is the generic part of what in practice may be a more extensive tool incorporating design and topic specific items.

We plan to do further work to develop these design and topic specific sections. We anticipate that certain items may need to be added for certain topic or design specific areas, while in other situations some of the items currently included in QUADAS may not be relevant and may need to be removed. Possible areas where the development of topic specific items may be considered include screening, clinical examination, biochemical tests, imaging evaluations, invasive procedures, questionnaire scales, pathology and genetic markers. Possible design specific areas include diagnostic case-control studies and diagnostic cohort studies. We plan to use a Delphi procedure to develop these sections. The Delphi panel will be larger than the panel used to develop the generic section and will include experts in each of the topic specific areas as well as experts in the methodology of diagnostic accuracy studies. Anyone interested in becoming a member of the Delphi panel for these sections can contact the authors.

One problem in using the QUADAS tool lies in the distinction between the reporting of a study and its quality. Inevitably the assessment of quality relates strongly to the reporting of results; a well conducted study will score poorly on a quality assessment if the methods and results are not reported in sufficient detail. The recent publication of the STARD document [14] may help to improve the quality of reporting of studies of diagnostic accuracy. The assessment of study quality in future papers should therefore not be limited by the poor quality of reporting which is currently a problem in this area. Currently, studies which fail to report on aspects of quality, for example if reviewers were aware of the results of the reference

standard when interpreting the results of the index test, are generally scored as not having met this quality item. This is often justified as faulty reporting generally reflects faulty methods. [15]

Another factor to consider when using QUADAS is the difference between bias and variability. Bias will limit the validity of the study results whereas variability may affect the generalisability of study results. QUADAS includes items which cover bias, variability and, to a certain extent, the quality of reporting. The majority of items included in QUADAS relate to bias (items 3, 4, 5, 6, 7, 10, 11, 12 and 14), with only two items each relating to variability (items 1 and 2) and reporting (items 8, 9 and 13).

Conclusions

This project has produced the first systematically developed evidence based quality assessment tool to be used in systematic reviews of diagnostic accuracy studies. Further work to validate the tool is in process.

Competing interests

None declared.

Authors' contributions

All authors contributed towards the conception and design of the study and read and approved the final manuscript. PW and AWSR participated in the performance of the Delphi procedure, the analysis and interpretation of data, and drafted the article. PMMB, JBR and JK provided supervision for the study and JK and PB obtained the funding for the study.

Additional material

Additional File 1

Table 1 – Initial list of items together with review evidence. Table 1 contains the list of 28 possible items for inclusion in the quality assessment tool which were rated using the Delphi procedure. It also contains the results of the systematic reviews on sources of bias and variation, and existing quality assessment tools.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-3-25-S1.doc]

Acknowledgements

We would like to thank Professor Colin Begg, Professor Patrick Bossuyt, Jon Deeks, Professor Constantine Gatsonis, Dr Khalid Khan, Dr Jeroen Lijmer, Mr David Moher, Professor Cynthia Mulrow, and Dr Gerben Ter Riet, for taking part in the Delphi procedure.

The work was commissioned and funded by the NHS R&D Health Technology Assessment Programme. The views expressed in this review are those of the authors and not necessarily those of the Standing Group, the Commissioning Group, or the Department of Health.

References

1. Glasziou P, Irwig L, Bain C, Colditz G: **Systematic reviews in health care: A practical guide.** *Cambridge: Cambridge University Press* 2001.
2. Deeks J: **Systematic reviews of evaluations of diagnostic and screening tests.** In: *Systematic Reviews in Health Care: Meta-analysis in context* Edited by: *Egger M, Davey Smith G, Altman D. London: BMJ Publishing Group*; 2001. Second edition
3. Whiting P, Rutjes A, Dinnes J, Reitsma JB, Bossuyt P, Kleijnen J: **A systematic review of existing quality assessment tools used to assess the quality of diagnostic research.** . submitted
4. Streiner DL, Norman GR: **Health measurement scales: a practical guide to their development and use.** *Oxford: Oxford University Press* 1995.
5. Jadad AR, Moore A, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ: **Assessing the quality of reports of randomised clinical trials: is blinding necessary?** *Control Clin Trials* 1996, **17:**1-12.
6. Juni P, Altman DG, Egger M: **Assessing the quality of controlled trials.** *BMJ* 2001, **323:**42-46.
7. Juni P, Witschi A, Bloch RM, Egger M: **The hazards of scoring the quality of clinical trials for meta-analysis.** *JAMA* 1999, **282:**1054-1060.
8. Greenland S: **Invited Commentary: A critical look at some popular meta-analytic methods.** *A J Epidemiol* 1994, **140:**290-296.
9. Whiting P, Dinnes J, Rutjes AWS, Reitsma JB, M BP, Kleijnen J: **A systematic review of how quality assessment has been handled in systematic reviews of diagnostic tests.** . submitted
10. Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J: **The development and validation of methods for assessing the quality and reporting of diagnostic studies.** *Health Technol Assess* in press.
11. Whiting P, Rutjes AWS, Reitsma JB, Glas A, Bossuyt PM, Kleijnen J: **A systematic review of sources of variation and bias in studies of diagnostic accuracy.** *Ann Intern Med, In press* .
12. Kerr M: **The Delphi Process.** *The Delphi Process 2002 City: Remote and Rural Areas Research Initiative, NHS in Scotland*; 2001. avaliable online at http://www.rararibids.org.uk/documents/bid79-delphi.htm
13. **The Delphi Technique in Pain Research.** *The Delphi Technique in Pain Research Volume 2002. City: Scottish Network for Chronic Pain Research*; 2001. avaliable at http://www.sncpr.org.uk/delphi.htm
14. Bossuyt P, Reitsma J, Bruns D, Gatsonis C, Glasziou P, Irwig L, Moher D, Rennie D, de Vet H, Lijmer J: **The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration.** *Clin Chem* 2003, **49:**7-18.
15. Schulz KF, Chalmers I, Hayes RJ, Altman DG: **Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials.** *JAMA* 1995, **273:**408-412.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/3/25/prepub

BMC Medical Research Methodology

BioMed Central

Research article

Open Access

Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies

Penny F Whiting*[1], Marie E Weswood[2], Anne WS Rutjes[3], Johannes B Reitsma[3], Patrick NM Bossuyt[3] and Jos Kleijnen[2]

Address: [1]MRC Health Services Research Collaboration, Department of Social Medicine, Canynge Hall, Whiteladies Road, Bristol, UK, [2]Centre for Reviews and Dissemination, University of York, UK and [3]Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, The Netherlands

Email: Penny F Whiting* - penny.whiting@bristol.ac.uk; Marie E Weswood - mew3@york.ac.uk; Anne WS Rutjes - a.rutjes@amc.uva.nl; Johannes B Reitsma - j.reitsma@amc.uva.nl; Patrick NM Bossuyt - p.m.bossuyt@amc.uva.nl; Jos Kleijnen - jos@kleijnen.freeserve.co.uk

* Corresponding author

Published: 06 March 2006

*BMC Medical Research Methodology*2006, **6**:9 doi:10.1186/1471-2288-6-9

Received: 01 September 2005
Accepted: 06 March 2006

This article is available from: http://www.biomedcentral.com/1471-2288/6/9

Abstract

Background: A quality assessment tool for diagnostic accuracy studies, named QUADAS, has recently been developed. Although QUADAS has been used in several systematic reviews, it has not been formally validated. The objective was to evaluate the validity and usefulness of QUADAS.

Methods: Three reviewers independently rated the quality of 30 studies using QUADAS. We assessed the proportion of agreements between each reviewer and the final consensus rating. This was done for all QUADAS items combined and for each individual item. Twenty reviewers who had used QUADAS in their reviews completed a short structured questionnaire on their experience of QUADAS.

Results: Over all items, the agreements between each reviewer and the final consensus rating were 91%, 90% and 85%. The results for individual QUADAS items varied between 50% and 100% with a median value of 90%. Items related to uninterpretable test results and withdrawals led to the most disagreements. The feedback on the content of the tool was generally positive with only small numbers of reviewers reporting problems with coverage, ease of use, clarity of instructions and validity.

Conclusion: Major modifications to the content of QUADAS itself are not necessary. The evaluation highlighted particular difficulties in scoring the items on uninterpretable results and withdrawals. Revised guidelines for scoring these items are proposed. It is essential that reviewers tailor guidelines for scoring items to their review, and ensure that all reviewers are clear on how to score studies. Reviewers should consider whether all QUADAS items are relevant to their review, and whether additional quality items should be assessed as part of their review.

Background

QUADAS is a tool to assess the quality of diagnostic accuracy studies included in systematic reviews. We defined quality as being concerned with both the internal and external validity of a study. QUADAS was developed in a systematic manner, based upon three reviews of existing

Table 1: Questionnaire for evaluation of QUADAS

a) Review details
b) Content of the tool:
• Did QUADAS cover all important items?
• Were any QUADAS items omitted, added or modified?
c) Background document:
• Was the background document easy to understand?
• Were scoring instructions understandable?
• Should any items have been scored differently?
d) Technical points
• How long did it take to complete QUADAS?
• Was inter-rater reliability assessed?
e) Overall conclusions
• Reviewers were asked to rate coverage, ease of use, clarity of instructions, and validity (whether QUADAS helped to distinguish between studies of different qualities) on a five point scale
• Would you us QUADAS again?
f) Additional questions
• How were the results of quality incorporated into the review?
• Was a training session organised to ensure reviewers applied the tool consistently?
• Reviewer details, including age, experience, professional background
• Have you previously been involved in the quality assessment of studies included in a systematic review?
g) Final comments

evidence and a Delphi procedure involving a panel of experts in diagnostic research [1]. Like all quality assessment tools, QUADAS is a measurement, implying that its characteristics have to be evaluated: does it measure what it aims to measure, how well does it do this, and are results reproducible between different observers [2]? The objective of this study was to evaluate QUADAS by determining agreement between reviewers and the consensus rating and variability among raters, and gathering feedback on reviewers' experiences of using QUADAS.

Methods

Assessment of the consistency and reliability of QUADAS

Three reviewers were asked to use QUADAS to independently rate the quality of 30 studies as part of a systematic review on the diagnosis of peripheral arterial disease. One QUADAS item, the use of an appropriate reference standard, was not assessed as studies were only included in the review if they used a specified reference standard.

The three reviewers had different backgrounds and levels of experience. Reviewer 1 had previously carried out several diagnostic systematic reviews and had used QUADAS; she also had a background in primary diagnostics. Reviewer 2 was a new reviewer – this was the first review that she had worked on, but she had previously worked in primary diagnostics. Reviewer 3 was an experienced reviewer who had worked on a number of systematic reviews. This combination of reviewers with was chosen to reflect the spectrum of likely QUADAS users.

A limited amount of information specific to the diagnosis of peripheral arterial disease was provided to help with the scoring of QUADAS, this applied to items 1 (spectrum composition), 4 (disease progression bias), and 12 (availability of clinical information). For all other items, the guidelines on scoring provided in the QUADAS background document were briefly summarised [3]. Although reviewers did have access to the background document they were not specifically requested to read this or use it when assessing study quality.

Our main interest was in the amount of agreement between the rating of each reviewer and the consensus rating, calculated as the proportion of studies for which each reviewer agreed with the consensus rating. In addition, we also examined inter-observer variability by calculating the kappa statistic. Both analyses were carried out for all QUADAS items combined and for each individual item. We chose to focus on the proportion of agreements between reviewers and the final consensus, as kappas can be misleading in certain circumstances [4].

Piloting QUADAS in ongoing reviews

Reviewers who had used QUADAS in their reviews completed a short structured questionnaire asking how they used QUADAS and what their opinions of its usefulness were. Details of the questionnaire are provided in Table 1. A narrative synthesis was used to summarise results.

Results

Assessment of the consistency and reliability of QUADAS

Table 2 summarises the agreement between reviewers. Agreement between reviewers 1 and 2 and the final consensus rating was very good at 91 and 90%, and was

Table 2: Overall agreement between reviewers and agreement with consensus for each of the QUADAS items and for all items combined

QUADAS item		Agreement with consensus diagnosis (%) (95% confidence interval)			Reviewer variability (κ) (95% confidence interval)
		1	2	3	
All items combined		91 (88–94)	90 (86–93)	85 (81–89)	0.66 (0.63 to 0.67)
1	Was the spectrum of patients representative of the patients who will receive the test in practice? (spectrum composition)*	90 (73–98)	87 (69–96)	83 (65–94)	0.73 (0.60 to 0.76)
2	Were selection criteria clearly described? (selection criteria)	90 (73–98)	83 (65–94)	73 (54–88)	0.55 (0.33 to 0.61)
3	*Is the reference standard likely to correctly classify the target condition? (reference standard)**				
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? (disease progression bias)*	87 (69–96)	90 (73–98)	83 (65–94)	0.68 (0.63 to 0.86)
5	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis? (partial verification)	87 (69–96)	90 (73–98)	93 (78–99)	0.27(-0.06 to 0.39)
6	Did patients receive the same reference standard regardless of the index test result? (differential verification)	97 (83–100)	97 (83–100)	97 (83–100)	0.31 (-0.01 to 0.46)
7	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)? (incorporation bias)	100 (88–100)	100 (88–100)	93 (78–99)	-0.02 (-0.03 to -0.01)
8	Was the execution of the index test described in sufficient detail to permit replication of the test? (index test execution)	97 (83–100)	100 (88–100)	87 (69–96)	0.60 (0.33 to 0.73)
9	Was the execution of the reference standard described in sufficient detail to permit its replication? (reference standard execution)	93 (78–99)	93 (78–99)	93 (78–99)	0.81 (0.60 to 0.87)
10	Were the index test results interpreted without knowledge of the results of the reference standard? (test review bias)	90 (73–98)	87 (69–96)	97 (83–100)	0.55 (-0.04 to 0.75)
11	Were the reference standard results interpreted without knowledge of the results of the index test? (reference standard review bias)	93 (78–99)	93 (78–99)	93 (78–99)	0.68 (0.46 to 0.76)
12	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? (clinical review bias)*	90 (73–98)	93 (78–99)	50 (31–69)	0.18 (-0.13 to 0.36)
13	Were uninterpretable/ intermediate test results reported? (uninterpretable test results)	83 (65–94)	70 (50–85)	87 (69–96)	0.32 (0.18 to 0.44)
14	Were withdrawals from the study explained? (withdrawals)	90 (73–98)	83 (65–94)	80 (61–92)	0.38 (0.33 to 0.51)

* Items for which review specific details were added to QUADAS

+The item relating to reference standard was not assessed for this review

slightly lower (85%) for reviewer 3. Overall reviewer variability was good [5] with a kappa of 0.65.

Agreement between reviewers and the final consensus rating was over 80% for all but four items: selection criteria, availability of clinical information, uninterpretable test results and withdrawals. The poor agreement for the availability of clinical information was related to reviewer 3 who had a very poor level of agreement (50%) with the final consensus rating; the other reviewers showed over 90% agreement with the final consensus. This suggests that reviewer 3 was interpreting this item differently to the other reviewers. The other three items, selection criteria, uninterpretable results and withdrawals, showed moderate agreement between each reviewer and the consensus rating suggesting that there may be difficulties in applying these items.

Piloting QUADAS in ongoing reviews

Twenty reviewers used QUADAS in their reviews and provided feedback via the structured questionnaire (Table 3). Fifteen reviewers came from the UK, two from Australia, two from the Netherlands, and one from Switzerland. Of those from the UK, seven were employees of the Centre for Reviews and Dissemination (CRD), which is where some of the researchers who developed QUADAS were based. The topics covered by the reviews included the diagnosis of: tuberculosis, urinary tract infection in children, haematuria, Dengue fever, prostate cancer, shoulder pain, epilepsy seizure focus, angina and myocardial infarction, infected diabetic foot ulcers, bacterial infections, lumbar fusion, multiple sclerosis, and osteoporosis. Diagnostic tests under evaluation included laboratory tests, imaging and physical examination. The number of studies included in the reviews ranged from 1 to 208 (median 28).

Content of tool

The feedback from 20 reviewers on the content of the tool was generally positive: eighteen reviewers thought that QUADAS covered all important items, seventeen did not omit any items, sixteen did not add any items, and nineteen did not modify any items.

Two reviewers thought that QUADAS did not cover all important items, one felt that it did not adequately cover population characteristics (description of spectrum, age, setting, prevalence), that questions regarding therapy, the positivity threshold of test results, and study design should have been included as separate items. These comments were mainly related to the desire to have information on these items so that they could be explored in subgroup analysis. The other reviewer thought that the tool should cover whether data could be extracted into a 2 × 2 table.

Three reviewers omitted items from QUADAS. One stated, "on occasions there were no withdrawals". One reviewer omitted items on: reference standard, disease progression bias, partial verification bias, differential verification bias and incorporation bias as these were not applicable to the topic area because there was no reference standard (the review was on prostate biopsies). The other reviewer omitted the item relating to disease progression bias as this did not apply to studies included in their review. Another reviewer stated that they did not omit any items but that as most of the studies included in their review were diagnostic case control studies, items on the availability of clinical information and withdrawals were difficult to answer, and in most cases the issue of follow-up was not relevant.

Four reviewers added items to QUADAS: one added clinically relevant items specific to their review, one added "Do you have plans to characterise data which are unsuitable for primary analysis?", one added "Was the raw data available?" and one added a number of items relating to the availability of 2 × 2 data, confidence intervals, a description of the index and reference tests and a description of the test threshold.

One reviewer modified the items on uninterpretable results and withdrawals to add a "not appropriate" response. She stated that if there were no uninterpretable test results it was unclear how to rate this item.

Background document

All but one reviewer found the background document easy to understand, two did not understand the scoring guidelines, and one reviewer thought that the items concerning differential and partial verification bias should have been scored differently. One reviewer found the item on disease progression bias difficult to understand. However, this difficulty appeared to be related to how to score this item specifically for their review rather than a problem with the instructions provided in the background document. Two reviewers stated that they added topic specific information to the background document to help determine exactly how to score items for their review.

Despite efforts to keep the wording of QUADAS simple to increase international applicability, two non-native English speakers had some difficulty in understanding the QUADAS background document. They found the item on the availability of clinical information difficult to understand and did not know what was meant by uninterpretable or indeterminate data or results, and felt that the background document did not clarify this. In future revisions, clarity of phrasing will be a key consideration.

The reviewer who thought items should have been scored differently felt that the items relating to verification bias should have been formulated differently and suggested "was verification bias avoided? (i.e. did the whole sample or a random selection of the sample receive verification using a reference standard)".

Technical points

The time taken to complete QUADAS ranged from less than 10 minutes to over an hour. Five reviewers reported that it took them <10 minutes, five that it took 10–15 minutes, seven that it took 15 to 30 minutes, two that it took 30 to 60 minutes and one that it took more than an hour. Some of the reviewers included the time to read the whole paper and carry out data extraction and completing QUADAS in this time, whereas others only included the time taken to complete QUADAS. None of the reviewers assessed inter-rater reliability.

Overall conclusions

Reviewers' ratings of QUADAS for coverage, ease of use, clarity of instructions and validity were generally good, especially for coverage, which was rated as good or very good by all reviewers, and ease of use, which was rated as at least average by all reviewers. One reviewer rated the clarity of instructions and the validity of QUADAS as being poor; she had earlier stated that she did not understand the instructions for scoring QUADAS. She also felt the studies in her review were of fairly poor quality but still fulfilled at least half the QUADAS items. All reviewers stated that they would use QUADAS again, although one stated that she may not use all 14 items next time and another stated that this was because there is currently no better tool available.

Additional comments

A major theme in reviewers' additional comments related to the poor quality of reporting of primary studies and the fact that this often limits the quality assessment. Another theme was that it is important to have an understanding of the clinical context while scoring some of the items. One reviewer suggested that it might be helpful to group the questionnaire using subheadings such as "general", "reference standard", and "index test". Another comment was that initial training on how to use the tool would be helpful.

Discussion

Principal findings

This evaluation has shown good agreement between reviewers and the final consensus rating for most QUADAS items and very positive feedback from reviewers who have used QUADAS. Two items, uninterpretable results and withdrawals, were found to be problematic. There was poorer agreement among reviewers and between reviewers and consensus for these items than for other items; feedback from reviewers also suggested problems with these items. One reviewer suggested that this might be because it is difficult to know what to do it if it is unclear if there are any uninterpretable results or withdrawals. Our own use of QUADAS supports this: we have found it very difficult to know how to score this item if the study does not report whether there were any uninterpretable results/withdrawals, and if all patients who entered the study appear to be accounted for. In such situations it is often unclear whether the study authors simply excluded uninterpretable results or withdrawals from their reports, or if there truly were no uninterpretable results or withdrawals. We have handled this problem by giving more explicit instructions for scoring these QUADAS items: we have stated that they should be scored as yes if it appears that all patients who were entered into the study completed the study.

The assessment of inter-rater reliability also highlighted possible problems with the items on the availability of clinical information and selection criteria. The item on clinical information is very specific to each review and it is therefore essential that clear guidelines on scoring this item be provided, outlining exactly what information should be available to the person interpreting the results of the index test. This definition should be agreed a priori. This was done for the review used for this evaluation and is reflected in the very high levels of agreement between two of the reviewers and the final consensus. It is unclear why the third reviewer showed much poorer agreement (50%) with the final consensus rating. It is unclear why the item on selection criteria showed poorer agreement with the consensus rating. This item was not highlighted as problematic in the feedback from reviewers. It may be related to the fact that no review specific information was provided for this item.

All additional items suggested for inclusion in QUADAS were considered as part of the development of QUADAS but were items that were not selected by the panel of experts for inclusion in the final tool. One of the items suggested for inclusion, the item relating to the threshold for the index test could be covered as part of item 8 (description of index test details). This is something to consider including in the guidelines for scoring this item when making guidelines specific to your review.

There was substantial variation in the time taken to complete QUADAS, ranging from less than 10 minutes to over 1 hour. This may be explained by the fact that some reviewers counted the time taken for the whole process of data extraction, including reading the paper, whereas others only counted the time taken to complete QUADAS. Despite this, half the reviewers took less than 15 minutes

Table 3: Summary of responses to the questionnaire on reviewers' experience of using QUADAS

Reviewer Number	Coverage?	Omit items?	Add items?	Modify item?	Easy to understand?	Scoring?	Different scoring?	Time to complete (minutes)	Inter-rater reliability ?	Coverage*	Ease of use*	Clarity of instruction*	Validity*	Use again?*
1	+	-	+	-	+	+	-	30–60	-	5	3	5	5	+
2	+	-	-	+	+	+	-	<10	-	4	4	3	4	+
3	+	+	-	-	+	+	-	<10	-	4	4	5	3	+
4	+	-	-	-	+	+	-	30–60	-	4	4	4	4	+
5	+	-	-	-	+	+	-	15–30	-	5	4	5	5	+
6	+	-	-	-	+	+	-	15–30	-	5	3	4	4	+
7	+	-	-	-	+	-	-	15–30	+	4	4	2	2	+
8	+	-	-	-	+	-	-	15–30	+	4	4	3	3	+
9	-	-	+	-	+	+	+	>60	-	4	5	5	4	+
10	+	-	-	-	+	+	-	<10	-	5	5	5	5	+
11	+	-	+	-	+	+	-	15–30	-	4	3	4	3	+
12	-	-	-	-	+	+	-	15–30	-	5	3	5	3	+
13	+	+	+	-	-	-	-	10–15	-	4	3	3	5	+
14	+	+	-	-	+	+	-	10–15	-	4	4	4	4	+
15	+	-	-	-	+	+	-	10–15	-	5	4	4	4	+
16	+	-	-	-	+	+	-	10–15	-	4	3	4	4	+
17	+	-	-	-	+	+	-	<10	-	4	4	4	4	+
18	+	-	-	-	+	+	-	<10	-	5	5	5	3	+
19	+	-	-	-	+	+	-	10–15	-	4	4	4	4	+
20	+	-	-	-	+	+	-	15–30	-	5	4	5	4	+

*Each of these items were rated from 1–5 where 1 is strongly disagree (very poor) and 5 is strongly agree (excellent)

Table 4: Proposed modifications to the QUADAS background document*

13. Were uninterpretable/intermediate test results reported?
c. How to score this item
If it is clear that all test results, including uninterpretable/indeterminate/intermediate are reported then this item should be scored as "yes". **If the authors do not report any uninterpretable/indeterminate/intermediate results, and if results are reported for all patients who were described as having been entered into the study then this item should also be scored as "yes".** If you think that such results occurred but have not been reported 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?
c. How to score this item
If it is clear what happened to all patients who entered the study, for example if a flow diagram of study participants is reported, then this item should be scored as "yes". **If the authors do not report any withdrawals and if results are available for all patients who were reported to have been entered into the study then this item should also be scored as "yes".** If it appears that some of the participants who entered the study did not complete the study, i.e. did not receive both the index test and reference standard, and these patients were not accounted for then this item should be scored as "no". If it is not clear whether all patients who entered the study were accounted for then this item should be scored as "unclear".
* Proposed changes are highlighted in bold

*Each of these items were rated from 1–5 where 1 is strongly disagree (very poor) and 5 is strongly agree (excellent)

and 17/20 took less than half an hour to complete QUADAS suggesting that QUADAS is relatively quick to complete.

Strengths and weaknesses of the study

The major strength of this study is that we carried out a detailed evaluation of QUADAS, which specifically included the views and experience of users. We are unaware of any other quality assessment tools for diagnostic accuracy studies that have undergone any process of evaluation.

Ideally, we would have liked to assess the "construct validity" of the tool – "the degree to which a test measures what it claims, or purports, to be measuring" [6]. As QUADAS aims to provide an indication of the quality of a study one way to assess this would be to take a set of "high" quality studies and a set of "low quality" studies and determine whether QUADAS can distinguish between these. This is known as "extreme groups" [6]. The problem with this process is determining which studies are high quality and which are low quality: there is no objective way of doing this. In addition, a systematic review is likely to include studies covering a range of quality. A quality assessment tool needs to be able to distinguish subtle differences across this full range of study quality, not just the extremes. We therefore decided against this method of evaluation.

Unanswered questions and future research

We originally proposed to carry out a meta-epidemiological regression analysis to investigate the association of individual QUADAS items with estimates of test performance. However, due to limited time and resources such an evaluation was not feasible. This is an area where future research would be beneficial. The Cochrane Collaboration is planning to extend its database to include diagnostic test accuracy reviews and is in the process of producing a handbook providing guidelines for the conduct of such reviews. The recommendations on quality assessment include a modified version of QUADAS (items 2, 8 and 9, the items relating to reporting rather than quality have been removed), and this will be built into the new Cochrane software. All diagnostic reviews included in the new Cochrane Database will therefore include an assessment of QUADAS with the results entered into the Review Manager Software in a structured way. In the future, once a number of Cochrane Test Accuracy Reviews have been completed, a meta-epidemiological regression analysis can be pursued.

Conclusions – Suggestions for modifications to QUADAS

We do not feel that major modifications to the content of QUADAS itself, in terms of items included, are necessary. However, the evaluation highlighted particular difficulties in scoring the items on uninterpretable results and withdrawals. We therefore recommend that the guidelines for scoring these items in the QUADAS background document be modified as shown in Table 4. In addition, we would like to highlight the importance of tailoring the guidelines for scoring items to each particular review, and of ensuring that all reviewers are clear on how studies should be scored for each of the items. It is not possible to provide a generic description of what should be considered an "appropriate patient spectrum", or what should be considered an "appropriate reference standard". It is therefore essential that all reviewers using QUADAS carefully consider how each individual item should be applied to their review and adapt the background document to make the guidelines for scoring specific to their review. This should be done in close collaboration with a clinical expert in the area of the review. Reviewers should also carefully consider whether all QUADAS items are relevant to their review, and also whether there are additional quality items not included in QUADAS which may

be of importance to their topic area and which they should assess as part of their review. Consensus should be established on all of these issues before starting the quality assessment. Lastly, an improvement in the quality of reporting, by endorsing the standards for reporting of diagnostic accuracy studies, the STARD initiative [7], should occur. This will allow reviewers to assess study quality rather than the quality of reporting.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

PW, JK and PB conceived the study. All authors contributed to the design of the study. PW and MW collected the data. PW carried out the analysis and drafted the paper. All authors commented on drafts of the manuscript and read and approved the final manuscript.

Acknowledgements

No financial or material support was provided for this study. We would like to thank the reviewers who participated in the assessment of inter-rater reliability and those who completed the questionnaire to provide feedback on their use of QUADAS.

References

1. Whiting P, Rutjes AWS, 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.
2. Bland JM, Altman DG: **Statistics Notes: Validating scales and indexes.** *BMJ* 2002, **324:**606-607.
3. Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J: **Development and validation of methods for assessing the quality of diagnostic accuracy studies.** *Health Technol Assess* 2004, **8:**1-234.
4. Lantz CA, Nebenzahl E: **Behavior and interpretation of the [kappa] statistic: Resolution of the two paradoxes.** *Journal of Clinical Epidemiology* 1996, **49:**431-434.
5. Altman DG: **14.3 Inter-rater agreement.** In *Practical Statistics for Medical Research* First edition edition. London, Chapman & Hall; 1999:403-408.
6. Brown JD: *Testing in language problems* Upper Saddle River, NJ, Prentice Hall Regents; 1996.
7. 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.** *Ann Intern Med* 2003, **138:**40-44.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/6/9/prepub