Updating QUADAS:

Evidence to inform the development of QUADAS-2

Penny Whiting, Anne Rutjes, Marie Westwood, Susan Mallett, Mariska Leeflang,

Hans Reitsma, Jon Deeks, Jonathan Sterne, Patrick Bossuyt

QUADAS Steering Group members:

Penny Whiting, Department of Social Medicine, University of Bristol

Anne Rutjes, Departments of Social and Preventive Medicine and Rheumatology, University of Berne

Jonathan Sterne, Department of Social Medicine, University of Bristol

Jon Deeks, Unit of Public Health, Epidemiology & Biostatistics, University of Birmingham

Mariska Leeflang, Department of Clinical Epidemiology, Biostatistics and Informatics, AMC, University of Amsterdam

Patrick Bossuyt, Department of Clinical Epidemiology, Biostatistics and Informatics, AMC, University of Amsterdam

Hans Reitsma, Department of Clinical Epidemiology, Biostatistics and Informatics, AMC, University of Amsterdam

Marie Westwood, Kleijnen Systematic Reviews, York

Susan Mallett, Centre for Statistics in Medicine, University of Oxford

Contents

Chapter 1: Background

QUADAS is a quality assessment tool for use in systematic reviews of diagnostic test accuracy (DTA) studies that we developed in 2003 (Table 1).(1)The steps we employed to develop QUADAS are outlined in Figure 1. We developed an initial list of possible items for inclusion through reviewing sources of bias and variation in DTA studies, reviewing existing quality assessment tools for DTA studies and examining how quality was incorporated into DTA reviews. We then conducted a Delphi procedure to refine this initial list of items to produce QUADAS. Members of the Delphi panel were experts in the area of diagnostic research. The process also included a preliminary evaluation of QUADAS which involved assessing inter-rater agreement in the rating of a set of 30 studies and gathering feedback from 20 reviewers who had used QUADAS in their reviews.(2) Based on this, some modifications were proposed for the scoring of two of the QUADAS items: interpretation of uninterpretable/intermediate test results and withdrawals.

Since its development QUADAS has been used in a large number of systematic reviews: it has been cited over 300 times and searching the DARE database using the term “QUADAS” identified 96 reviews. A modified version of QUADAS, with items related to the quality of reporting removed, has been adopted for use by the Cochrane Collaboration and is recommended for use in all Cochrane DTA reviews.(3) QUADAS has also been recommended for use by NICE. Our own experience, anecdotal reports, and feedback via Cochrane suggest some problems with the current version of QUADAS. These include problems in scoring certain items (in particular items on spectrum, uninterpretable/intermediate test results and withdrawals), possible overlap between items (for example partial verification bias and withdrawals), and certain situations where it is difficult to use QUADAS (for example in topics in which the reference standard involves an element of follow-up). We therefore decided to revisit QUADAS with the aim of using the experience gathered through its use and new evidence regarding sources of bias and variation to update QUADAS to produce “QUADAS-2”.

Figure 1: Development of original QUADAS tool

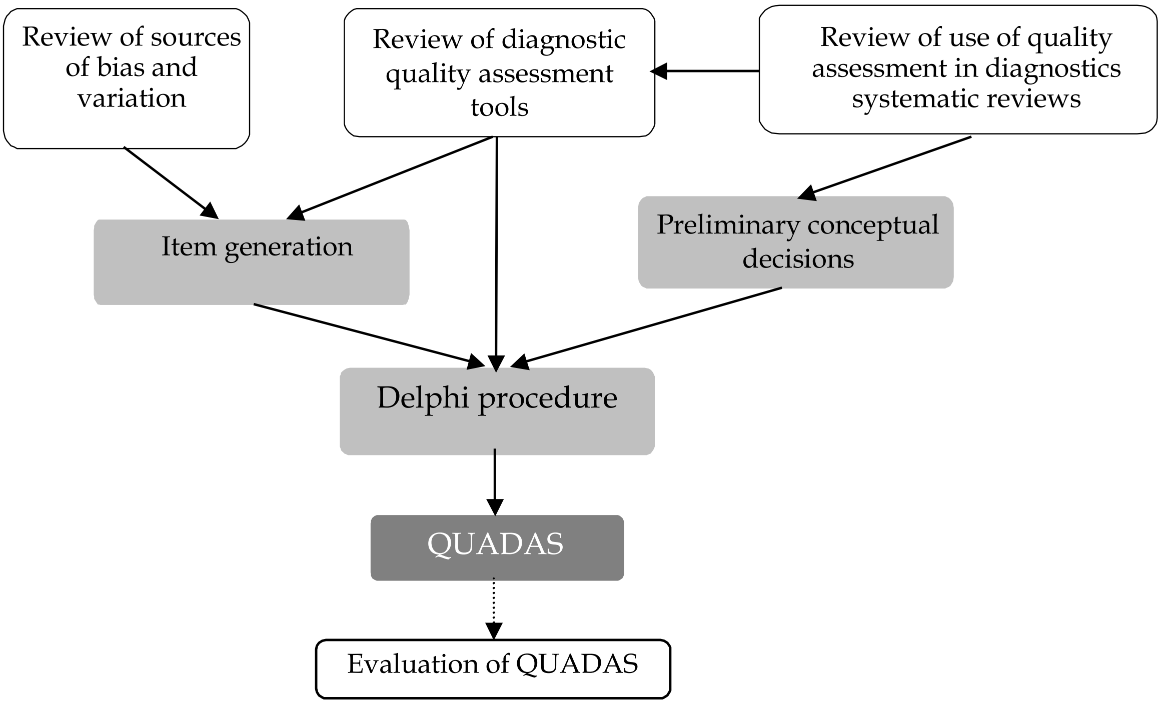

Review of sources of bias and variation
Review of diagnostic quality assessment tools
Review of use of quality assessment in diagnostics systematic reviews
Item generation
Preliminary conceptual decisions
Delphi procedure
QUADAS
Evaluation of QUADAS

Table 1: 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?			

Items in italics are those removed from the Cochrane version of QUADAS.

Chapter 2: Approach and Scope of QUADAS-2

Key points

Experience and feedback suggest that QUADAS needs revision.

We suggest adapting an approach proposed by Moher for guideline development, including a face-to-face meeting, to develop QUADAS-2

We have used a four-phase approach to inform the development of QUADAS-2:

1. Review of the use of QUADAS (Ch3)
2. Formal feedback from reviewers (Ch 4)
3. Review of new evidence on sources of bias and variation (Ch5)
4. Review of studies that have performed an evaluation of QUADAS (Ch6)

Conceptual decisions

- QUADAS-2 will have the same general requirements as the original QUADAS tool
- Change scoring from “yes/no/unclear” to “low/high/unclear risk of bias”
- Similar structure to Cochrane risk of bias tool
- Sub-items will be added to facilitate scoring of, for example, partial verification
- Topic specific items or items concerning prognostic studies will not be added
- Item(s) addressing comparative designs and those including follow-up will be added
- Striving for holistic tool, avoiding overlap between items

We have selected the approach proposed by Moher et al.(4) to develop QUADAS-2 (Table 2). Although this approach was proposed for guideline development, most of the proposed stages apply equally to the development of a quality assessment tool. The main focus will be a face-to-face consensus group meeting of experts in the area of diagnosis. This report summarises the results of the pre-meeting activities, in particular, the rationale and scope of QUADAS-2 and the evidence base for the development of QUADAS-2. Separate summary documents will be developed for the “face-to-face consensus meeting”, “post-meeting activities” and “post-publication activities”.

Table 2: Proposed stages for the development of QUADAS-2: adapted from Moher et al. “Reporting Guidance to Developers of Health Reporting Guidelines"

Pre-meeting activities	
Item #	
1	Funding the guideline initiative
2	Rationale and scope of QUADAS-2
3	Develop the evidence base - Review on how quality has been assessed and incorporated into DTA reviews - Feedback from reviewers who have used QUADAS - Update SR on sources of variation and bias in DTA studies - Review of studies that evaluated QUADAS
4	Generating a list of items for consideration
5	Organization and logistics of QUADAS-2 development
5a	Identify group members
5b	Decide size and duration of the meeting
5c	Book the meeting venue
5d	Develop meeting logistics
5e	Develop meeting agenda
5f	Prepare materials to be sent to participants prior to meeting
5g	Arrange to record the meeting
Face-to-face consensus meeting activities	
7[†]	Present and discuss results of pre-meeting activities and relevant evidence
8[†]	Discuss the rationale for including items in the checklist
9[†]	Generate items for inclusion in checklist
11[†]	Discuss strategy for producing documents; identify who will be involved in which activities; discuss authorship
12	Discuss knowledge translation strategy
Post-meeting activities	
13[†]	Develop QUADAS-2
14	Pilot QUADAS-2
15	Develop background document
16	Develop a publication strategy

Post-publication activities	
18[†]	Seeking and dealing with feedback and criticism
20	Website development?

2.1 Rationale for QUADAS-2

It has been almost 10 years since the development of the original QUADAS tool. Over this time it has been used in numerous reviews and a wealth of information on its use is available. Experience through using QUADAS and feedback received from reviewers who have used QUADAS, in particular Cochrane reviewers, have highlighted some problems with using QUADAS. We have therefore decided to revisit QUADAS with the aim of using the experience gathered through its use and new evidence regarding sources of bias and variation to update QUADAS to produce “QUADAS-2”.

2.2 Scope of QUADAS-2

The following decisions regarding the scope of QUADAS-2 were made by the steering group.

Preliminary conceptual decisions

QUADAS-2 will have the same general requirements as the original QUADAS tool:

- Be used in systematic reviews of DTA studies
- Assess the methodological quality of a DTA study in generic terms (relevant to all DTA studies)
- Allow consistent and reliable assessment of quality by reviewers with different backgrounds
- Be able to distinguish between high and low quality studies
- Be relatively short and simple to complete
- Should not incorporate a quality score

Definition of quality

The definition of quality used for the original QUADAS tool was:

“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.”

In practice we also extended this definition to include quality of reporting. However, when we modified QUADAS for use by Cochrane the items relating to the quality of reporting were removed.

For the development of QUADAS-2, we propose following the terminology used for Cochrane reviews of interventions and moving away from the term quality and instead using the phrase "risk of bias". We therefore suggest the following revised definition of quality:

"both the risk of bias and applicability of a study; (1) the degree to which estimates of diagnostic accuracy avoided risk of bias, and (2) the extent to which primary studies are applicable to the review's research question"

One of the major changes for QUADAS-2 that we propose is to restructure the tool to include two separate sections, one focusing on risk of bias and the other on applicability. We will not include items relating to quality of reporting.

Scoring

QUADAS currently consists of a series of questions each of which is rated as "yes", "no", or "unclear", where yes always indicates an absence of bias. This simple method of scoring has generally received positive feedback from reviewers. The "Risk of Bias" tool developed for Cochrane reviews of interventions has moved away from this method of scoring to a rating of "high risk of bias" or "low risk of bias". We suggest that the scoring of the risk of bias component of QUADAS-2 follows this approach. The scoring will also follow the Cochrane structure of separating the description of the basis for the scoring from the judgement of risk of bias. We need to further consider how this can be adapted for the section of QUADAS-2 relating to applicability.

Sub-items

We will expand QUADAS so that in addition to the key items, which we will aim to limit to as few as possible, we will add sub-items which will help to allow objective assessment of the key items. For example, scoring partial verification needs to generate data on the number of non-verified patients, the pattern (for instance index test negatives only or T+ and T-), and how these patients were handled in the analyses (removed, classified as TN or TP, or correction method, or imputation). This information is needed to judge the potential direction and size of the bias. Inclusion of sub-items means that such items could be assessed individually before providing the overall assessment of partial differential verification bias.

Comparative tests

The current version of QUADAS does not cover the situation of comparative tests. This is a limitation as more and more reviews are covering topics which include comparison of multiple tests. We will aim to cover the situation of comparative tests in QUADAS-2.

Prognostic/predictive tests

We considered extending QUADAS-2 to cover index tests used for prognosis and/or prediction, but decided that this was not feasible within a single quality assessment tool and that such a situation needs to be covered in a separate tool and as such is beyond the remit of QUADAS-2.

Topics that involve some degree of longitudinal follow-up

The classic DTA study applies an index test to all patients suspected of having the target condition and then applies the reference standard to these patients at approximately the same point in time, and so is essentially cross-sectional in design. However, there are many situations in which the reference standard involves some degree of follow-up. For example, a firm clinical diagnosis of multiple sclerosis (MS) can only be made several years after the patient initially presents with possible symptoms. Another common situation is diagnosis in pregnancy where tests are applied during pregnancy but the diagnosis is not confirmed until after the birth. Many screening tests can be applied in pregnancy Studies evaluating new tests for the early diagnosis of MS have to incorporate a degree of follow-up in the reference standard. QUADAS does not currently take this into consideration. We propose that this is covered by QUADAS-2.

Similar structure to Cochrane risk of bias tool

The Cochrane risk of bias tool for use in reviews of interventions includes a section to collect details on the basis on which the scoring was made. For example, for randomisation in addition to scoring the study according to whether or not the method of randomisation was appropriate, details are also extracted on methods used to randomise patients. The application of QUADAS in RefMan-5 already follows this approach with the inclusion of fields to explain the scoring for each QUADAS item. We suggest adopting this structure for QUADAS-2.

Holistic nature of the tool

When developing QUADAS-2 we need to aim to develop a set of independent criteria that work together, i.e. to ensure that there is no overlap between items. This was a major consideration when we decided to have a face-to-face meeting as the main activity for developing QUADAS-2, rather than using a Delphi-procedure as used when developing the original QUADAS tool.

Topic specific items

We considered broadening the scope of QUADAS to include topic-specific items either for test type (e.g. imaging, biochemistry), or clinical field. We decided not to expand QUADAS-2 to include topic specific items but will keep an additional list of possible items.

2.2 Develop the evidence base

We used a four phase approach to provide the evidence base to inform the development of QUADAS-2. The results of each of these phases are summarised in this report.

Phase 1: Overview of how study quality has been assessed and incorporated into DTA reviews, with a particular focus on the use of QUADAS (Chapter 3)

We examined 54 DTA reviews, half of which were selected on the basis of having used QUADAS, to investigate how quality was assessed and incorporated into a sample of recent DTA reviews. The information provided from this review was used to evaluate how QUADAS has been used in practice, highlight items which may be problematic, and to identify items for possible inclusion/exclusion for QUADAS-2.

Phase 2: Feedback from Reviewers (Chapter 4)

QUADAS has been available for use in DTA reviews since 2003 and since then has been used in a large number of reviews covering a variety of topics. We developed a simple web-based questionnaire, to gather structured feedback from reviewers who have used QUADAS. We invited all first authors of reviews indexed on DARE that have used QUADAS (96 reviews) and authors of protocols and completed Cochrane DTA reviews to complete the questionnaire. We also encouraged all invitees to circulate details of the questionnaire to other reviewers who may have used QUADAS.

Phase 3: Update review on sources of bias and variation (Chapter 5)

We have updated our review on sources of bias and variation in DTA studies.(5) Searches for the original review were conducted to September 2001; these were updated to cover the intervening period (2001-2010). We have updated the results of the original review to incorporate 46 additional studies.

Phase 4: Review of studies that have evaluated QUADAS (Chapter 6)

A number of studies have been published reporting on reviewer's experience of using QUADAS and of inter-rater reliability. We identified 8 studies that had reported such evaluations.

2.3 Generate a list of items for consideration for inclusion in QUADAS-2 (Chapter 7)

Based on the results of the four phases of evaluation of QUADAS, we identified original QUADAS items to be retained in QUADAS-2, items that are problematic and need reworking for QUADAS-2, items to be removed, and possible new items for inclusion.

Chapter 3: Diagnostic test accuracy (DTA) reviews: Conduct and reporting of quality assessment

Key points

54 DTA reviews were included in the evaluation of the quality of conduct and use of quality assessment

33 reviews used QUADAS:

- Item on Patient spectrum was kept in all QUADAS and in near all other reviews
- Patient spectrum was modified in some reviews by involving additional subcategories usually relating to study design.
- Most commonly omitted items were the availability of clinical information (item 12), avoidance of incorporation bias (item 7) and use of an appropriate reference standard (item 3); reviews that omitted the item relating to reference standard generally restricted inclusion based on reference standard
- Although some reviews added additional quality items there were no items that were consistently added across multiple reviews
- Quality scores were used in one third of the reviews

21 reviews did not use QUADAS

- 8 did not perform a formal quality assessment
- Item on Patient spectrum was used in nearly all reviews
- Less items were covered in the quality assessment
- Quality scores were used in two thirds of the reviews

This review aims to provide an overview on how study quality has been assessed and incorporated into DTA reviews, with a particular focus on the use of QUADAS. The results of this review will be used to inform the development of QUADAS 2.

3.1 Objectives

- To review the quality of conduct and reporting of quality assessment in a sample of DTA reviews
- To evaluate how QUADAS has been used in published DTA reviews
- To inform the development of QUADAS2

3.2 Methods

We searched the DARE online database using the term "QUADAS" to identify DTA reviews that have used QUADAS ("QUADAS reviews"). We obtained a list of all full and provisional DTA abstracts on DARE from CRD, deduplicated against the QUADAS reviews to give a list of DTA reviews that did not use QUADAS ("DARE reviews"). The review was restricted to reviews considered to be true DTA reviews – those that assess the results of an index test in comparison to a reference standard and report cross tabulation of results.

We grouped reviews according to the following topic areas: clinical, biochemical, histology, imaging, questionnaire, other, and combination across categories. We selected the five most recent reviews from each category for DARE reviews and QUADAS reviews. When multiple reviews in a single category were published in a single year, we used a random number generator to randomly select the appropriate number of reviews from within that year. If less than five reviews were available for a single category then all reviews in this category were selected. We aimed to include a minimum of 50 reviews.

We developed a data extraction form using MS Access to collect data from the included reviews (Appendix 1). This included all items relating to how quality was incorporated into the review assessed in our previous review on this topic.(6) This allowed assessment of whether uptake of QUADAS has had any influence on how quality is assessed and incorporated into DTA reviews. One reviewer performed the data extraction; this was checked by a second reviewer.

We categorised reviews according to review topic and use of QUADAS in order to investigate whether methods used for quality assessment in DTA reviews differs according to review topic and use of QUADAS. We recorded when reviews assessed diagnostic tests composed of multi-component scores such as patient questionnaires, because there are additional quality issues when reviews include a mixture of articles deriving new scores and articles externally validating scores. To investigate how quality assessment in DTA reviews has changed over time, we compared the results from this review to the findings of our previous review on how quality is incorporated into DTA reviews.(6)

3.3 Results

General Details

We included 54 DTA reviews. Details of each review are provided in Appendix Table 3.1 and are summarised in Table 3.1. Each of the following categories were assessed by eight reviews: biochemical, clinical, test combinations, imaging and other. Six reviews assessed histological tests. In at least three reviews the diagnostic test was a patient questionnaire used to form a multi-component score, potentially containing additional sources of bias to other diagnostic tests. These reviews included a mixture of articles which were external validation studies and articles with additional high bias as the results were from the same population that the multi-component score was developed in. One review included a quality item to capture the additional bias in some of the included studies, by assessing whether the study evaluated test performance in a population other than that used to derive the multi-component instrument. The reviews using these multi-component scores were categorised as questionnaires in Table 3.1.

All reviews defined inclusion criteria in terms of the index test and most (94%) defined inclusion criteria in terms of the target condition. Around 70% of reviews defined inclusion in terms of population, reference standard and outcomes (e.g. 2 x 2 data) but only 60% specified inclusion criteria in terms of study design. Although 60% of reviews defined the proposed role of the index only 35% restricted inclusion to studies that assessed the test in this role and 43% of reviews restricted inclusion to patients in whom the test will be used in practice. The majority of studies conducted a formal quality assessment and just over 60% of reviews used QUADAS to assess study quality. A further two reviews reported that QUADAS had been used but referenced other publications and did not use criteria related to QUADAS. These reviews were considered not to have used QUADAS.

Table 3.1: Summary of included DTA reviews

Topic	N	Were inclusion criteria defined in terms of:						Was Index test role defined?	Was inclusion Restricted to studies of this role?	Was Inclusion restricted to patients in whom	Was a QA conducted?	Was QUADAS used?
		Population	Index test	Target condition	Reference standard	Outcome	Study design					
Biochemical	8	3 (38)	8 (100)	8(100)	5(63)	5 (63)	4(50)	4(50)	1(13)	2(25)	7(88)	5(63)
Clinical	8	6 (75)	8(100)	8(100)	6(75)	5 (63)	5(63)	5(63)	3(38)	5(63)	7(88)	5(63)
Combination	8	6 (75)	8(100)	7 (88)	7(88)	6(75)	5(63)	5(63)	3(38)	1(13)	7(88)	5(63)
Histology	6	6 100)	6(100)	5 (83)	4(67)	4(67)	4(67)	4(67)	3(50)	4(67)	3(50)	2(33)
Imaging	8	5 (63)	8(100)	8(100)	7(88)	7(88)	5(63)	7(88)	4(50)	4(50)	8(100)	6(75)
Other	8	6 (75)	8(100)	7(88)	4(50)	6(75)	5(63)	3(38)	2(25)	2(25)	7(88)	6(75)
Questionnaire	8	6 (75)	8(100)	8(100)	6(75)	5(63)	3(38)	6(75)	3(38)	5(63)	7(88)	4(50)
Total	54	38 (70)	54(100)	51(94)	39(72)	38(70)	31(57)	34(63)	19(35)	23(43)	46(85)	33(61)

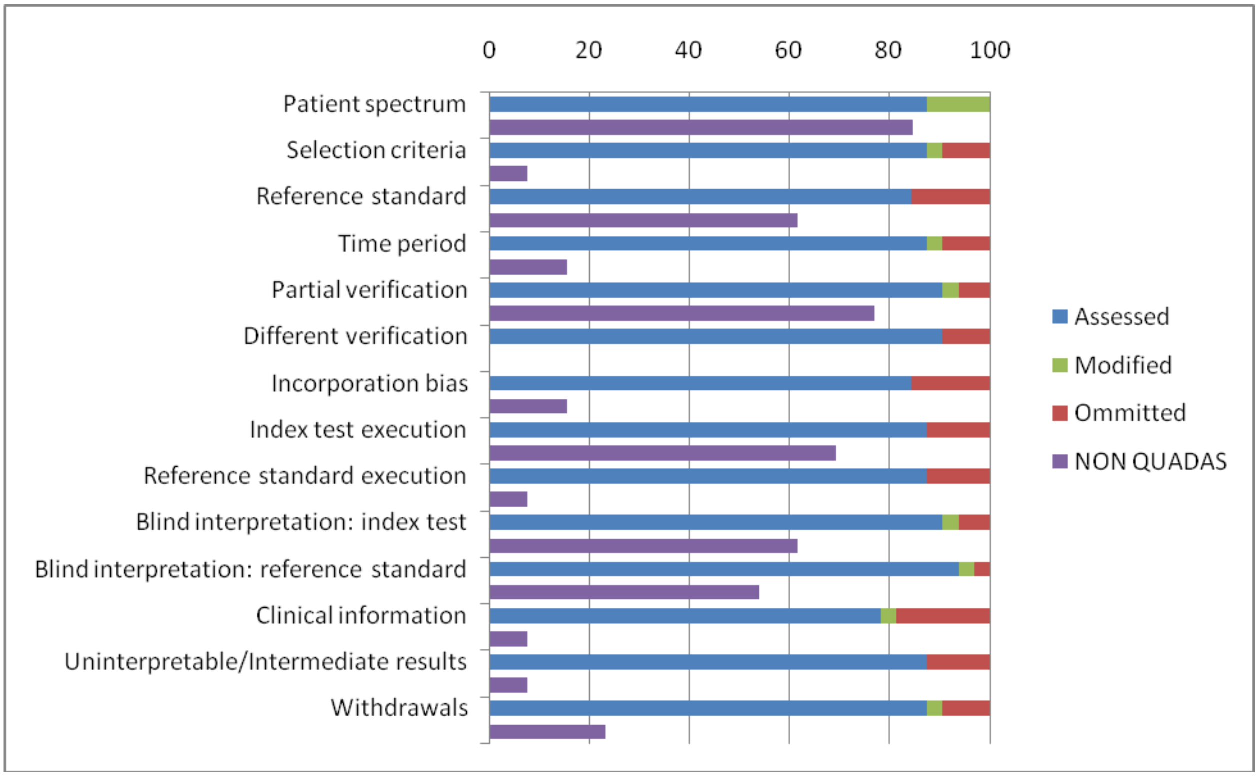
Assessment of study quality using QUADAS

Of the 33 reviews that used QUADAS to formally assess study quality, 20 (61%) used QUADAS without modification. One review did not provide sufficient details to judge which QUADAS items were considered.{ref} Table 3.2 summarises the number of reviews that assessed, omitted or modified each QUADAS item and Appendix Table 3.2a provides a more detailed overview of the items assessed by these reviews. None of the QUADAS items were assessed by all reviews that used QUADAS. The items relating to blinding (items 10 & 11) were each assessed by over 90% of reviews. Four reviews modified item 1 (patient spectrum). Modifications generally involved splitting this item into additional subcategories such as study design and sampling method. Items 3 (reporting of selection criteria), 5 (partial verification bias), 10 (blinding of index test to reference standard results), 11 (blinding of reference standard to index test results) and 14 (withdrawals) were each modified in single reviews; other items were not modified in any reviews.

Omission of items occurred more frequently than modification although reasons for omission were not always reported. The only item not to be omitted by any reviews was item 1. The most frequently omitted item was item 12 (clinical review bias) which was omitted by six reviews. Reasons for omission were reported in four of these and included it not being relevant as the review was evaluating clinical criteria, the review was evaluating automated tests and no interpretation was involved, unclear what clinical information was available in the primary studies and could not be operationalised for the studies included in the review. The use of an appropriate reference standard was also frequently omitted (5 studies), three reviews reported that this was because inclusion was restricted based on reference standard the other two reviews did not report reasons for omissions.

Table 3.2/Figure 3.1: Number of reviews that assessed, omitted or modified each QUADAS item and number of reviews that did not use QUADAS but that assessed equivalent items

Item		Assessed (%)	Omitted	Modified	Non-QUADAS reviews (%)
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	28 (88)	0	4	11 (85)
2.	Were selection criteria clearly described?	28 (88)	3	1	1 (8)
3.	Is the reference standard likely to correctly classify the target condition?	27 (84)	5	0	8 (62)
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?	28 (88)	3	1	2 (15)
5.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	29 (91)	2	1	10 (77)
6.	Did patients receive the same reference standard regardless of the index test result?	29 (91)	3	0	0
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	27 (84)	5	0	2 (15)
8.	Was the execution of the index test described in sufficient detail to permit replication of the test?	28 (88)	4	0	9 (69)
9.	Was the execution of the reference standard described in sufficient detail to permit its replication?	28 (88)	4	0	1 (8)
10.	Were the index test results interpreted without knowledge of the results of the reference standard?	29 (91)	2	1	8 (62)
11.	Were the reference standard results interpreted without knowledge of the results of the index test?	30 (94)	1	1	7 (54)
12.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	25 (78)	6	0	1 (8)
13.	Were uninterpretable/ intermediate test results reported?	28 (88)	4	0	1 (8)
14.	Were withdrawals from the study explained?	28 (88)	3	1	3 (23)



Six reviews used alternative criteria (STARD{503/id}, AHRQ (7) and Deville(8)) in addition to QUADAS, and a further six reviews added additional items to the quality assessment. These included items covering funding (2 reviews), determination of threshold prior to study commencement (2 reviews), prospective recruitment (1 review), proportion of patients recruited enrolled (1 review), inter-observer variability (1 review), evaluation of current test technology (1 review), reporting of definition of positive test result prior (1 review), administering of preventive intervention (1 review), whether the results were valid (1 review), description of setting for the test interpretation (1 review), patient or segment unit of analysis (1 review), and reporting of methods of analysis (1 review). One review of a multi-component diagnostic score derived from a patient questionnaire included a quality item to capture the additional bias in some of the included studies, by assessing whether the study evaluated test performance in a population other than that used to derive the multi-component instrument.

In addition to the formal quality assessment undertaken thirteen studies also incorporated quality into the review using more informal processes. Eight reviews did this by restricting inclusion based on the following: single defined reference standard (5 reviews), study design (e.g. prospective enrolment, exclusion of case-control studies; 4 reviews) and reporting of data for at least 50% of patients enrolled (1 review). Four reviews extracted data relating to study quality (study design/enrolment: 4 reviews; previous test results: 1 review; sample size: 1 review; observer variability: 1 review) in addition to the formal quality assessment and a further review investigated items relating to study design and enrolment as possible sources of heterogeneity.

Assessment of study quality in reviews that did not use QUADAS

Twenty one of the included reviews did not use QUADAS. Of these, eight did not conduct a formal quality assessment although three of these did use informal process to incorporate quality into the review. (9) (10;11) All three restricted inclusion based on a single reference standard and two also investigated quality related features (% insufficient material, study design and blinding) as possible sources of heterogeneity.

Two reviews used published criteria (Sackett criteria(12), CASP programme(13)) to assess study quality and a further seven reviews adapted existing criteria: US Preventive Services Task Force criteria(14) (3 reviews), CRD Report 4 (2001)(15) (1 review), Honest (2002)(16) (1 review), Kelly et al.(17)(1 review), and Lijmer (1999)(18) (1 review). The remaining four studies used criteria developed by the authors for the review. Appendix Table 3.2b summarises details of the items assessed by the reviews that did not use QUADAS and maps the items assessed to their equivalent QUADAS Item. Table 3.3 summarises the number of reviews that assessed items equivalent to each of the QUADAS items. All but two of the reviews included items related to item 1 (Patient spectrum) with the majority of these including multiple items such as whether studies were prospective, whether recruitment was consecutive and details relating to the enrolled participants. None of the studies explicitly assessed differential verification bias and items 2 (reporting of selection criteria), 9 (execution of the reference standard), 12 (availability of clinical information) and 13 (reporting of uninterpretable/intermediate results) were each assessed in single reviews.

Five reviews used additional informal methods of incorporating quality by restricting inclusion based on reference standard (2 reviews), appropriate study design (2 reviews), prospective enrolment (1 review), avoidance of disease progression bias (1 review), and avoidance of partial verification bias (1 review) and by investigating the presence of heterogeneity based on whether studies avoided disease progression bias (1 review).

Scoring methods used

Of the reviews that used QUADAS, ten (30%) explicitly described scoring guidelines for at least one QUADAS item, modified specifically for the review. Reviews generally followed the recommend method of rating QUADAS items as "Yes/No/Unclear" with twenty reviews (61%) using this exactly, one review modified this slightly to "Yes/No/Not reported", two reviews added a "not applicable" category, one review rated items as "yes/no/can't tell" and for some items added additional descriptive categories. In three reviews it was unclear how items were rated. One of the reviews that did not use QUADAS used the "Yes/No/Unclear" rating approach recommended by QUADAS. Four reviews used slight variations on this scoring ("Yes/No/Not available"; "Yes/No/Not reported"; 1 if criterion met, 0 if no or unclear; +/ -/ +- (partially fulfilled)) and a further review simply rated items as "Yes or No". One review used descriptive categories to summarise the results of the quality assessment and in the remaining six reviews it was unclear how items were rated.

Use of quality scores

Despite specific recommendations within the guidelines that accompany QUADAS, almost one third of the reviews (11) that used QUADAS estimated summary quality scores based on the QUADAS assessment. This was generally done by simply summing the number of items fulfilled to give a score out of 14. Almost half of the reviews that did not use QUADAS reported summary quality scores in the review.

Grouping of studies based on quality

Eleven (33%) of the reviews that used QUADAS stratified studies according to quality. Six of the reviews classified studies as high or low quality based on achieving a summary quality score above a specified value – four reviews used a cut-off of 10/14, one used 11/14, and one used the median quality score. A further review classified studies as high, moderate, low, or very low quality based on summary quality scores. One review classified studies as being of low quality if they "failed" (i.e. scored no) 3 or more QUADAS items. Two reviews defined key quality items and considered studies to be of high quality if all or a pre-specified number of these were fulfilled. One review considered studies to be of high quality if they enrolled an appropriate patient spectrum (scored yes for QUADAS item1).

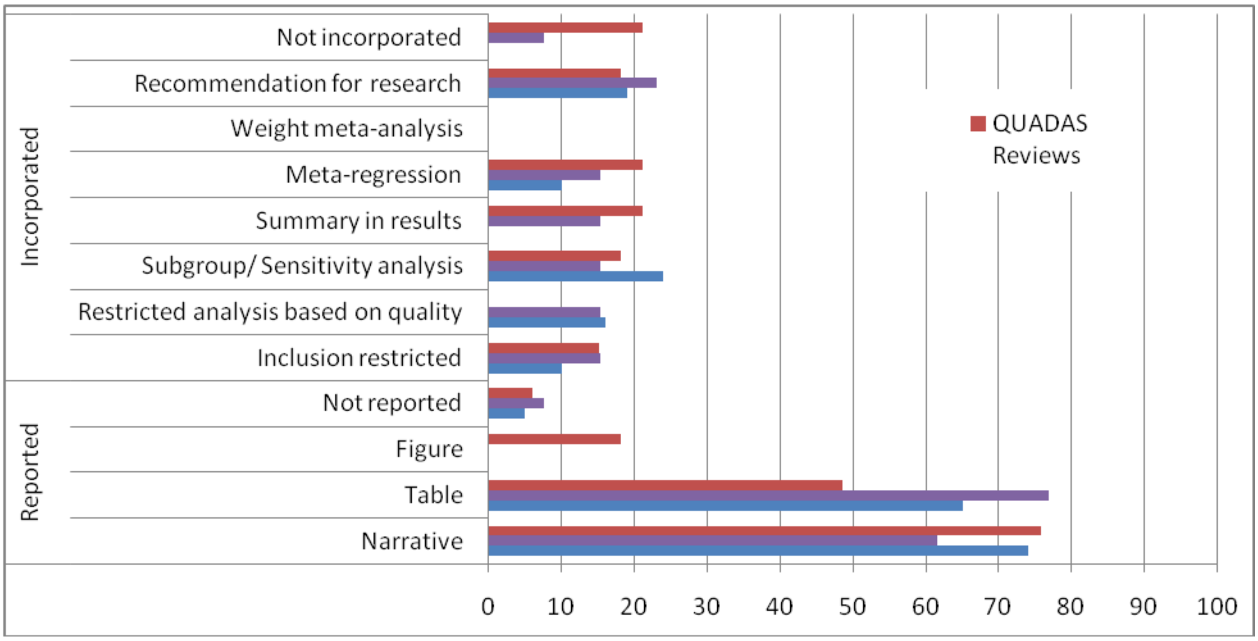
Nine of the reviews that did not use QUADAS stratified studies according to quality. Two reviews classed studies as high quality based on summary quality scores and a further four reviews grouped studies into multiple categories based on summary scores. One review defined high quality studies as those fulfilling specific criteria and one review classed studies as high quality if they fulfilled all five quality criteria assessed. The final study stated that poor quality studies were included but lacked details on how these were defined.

Incorporation of quality into the review

Appendix Table 3.2b provides full details of how each review reported the results of the quality assessment and incorporated the quality assessment into the review. Table 3.3/Figure 3.2 provides a summary of this information across reviews.

Table 3.3/Figure 3.2: Details on how quality was reported and incorporated in the review

Method of reporting/incorporating study quality		QUADAS Reviews n(%)	Non-QUADAS reviews n (%)	2004 Review n(%)
How were the results of the QA reported?	Narrative	25 (76)	8 (62)	43 (74)
	Table	16 (48)	10 (77)	38 (65)
	Figure	6 (18)	0	Not assessed
	Not reported	2 (6)	1 (8)	3 (5)
How were the results of the QA incorporated?	Inclusion in review	5 (15)	2 (15)	6 (10)
	Inclusion in primary analysis	0	2 (15)	9 (16)
	Subgroup/ Sensitivity analyses	6 (18)	2 (15)	14 (24)
	Summary in results	7 (21)	2 (15)	Not assessed
	Meta-regression	7 (21)	2 (15)	6 (10)
	Weight meta-analysis	0	0	Not assessed
	Recommendation for research	6 (18)	3 (23)	11 (19)
	Not incorporated	7 (21)	1 (8)	Not assessed



The majority of reviews, both those that used QUADAS and non-QUADAS reviews, provided at least a narrative summary of the results of the quality assessment or reported results in a Table, or for QUADAS reviews, a Figure. Six reviews reported summary quality scores in Tables of study details. Two of the reviews that used QUADAS and one of the non-QUADAS reviews did not provide any details of the results of the quality assessment.

Over 20% of the QUADAS reviews did not incorporate the results of the quality assessment into the review synthesis; only one of the non-QUADAS reviews did not incorporate quality. Methods used to incorporate quality into the review included presenting a narrative summary in the results relating the results of the studies to items included in the quality assessment, restriction of the review or primary analysis based on quality, subgroup/sensitivity analysis or meta-regression to investigate the association of various quality items with measures of accuracy and as a basis for recommendations for future research. Each of these methods were used by around 15 to 20% of reviews with similar proportions for both QUADAS and non-QUADAS reviews. None of the reviews used quality to weight the meta- analysis.

Comparison with previous review

The results of the current review were similar to that of the review that we conducted in 2005.(6) The only apparent difference was the slightly smaller proportion of reviews in the 2005 review that used meta-regression to investigate the association of quality items with measures of accuracy (10% vs. 21% and 15%) and the larger number of reviews using subgroup/sensitivity analysis to investigate the association of quality items with measures of accuracy (24% vs. 15% and 18%).

Specific problems with QUADAS reported by the included reviews

A number of reviews highlighted specific problems associated with using QUADAS. These included poor reporting in primary studies in particular in relation to index test and reference standard execution, uninterpretable results, withdrawals and availability of clinical information. One review reported that most disagreements related to use of the same reference standard (item 6) and incorporation bias (item 7). This review highlighted the importance of including review specific guidelines for scoring.

3.4 Summary

Quality criteria

Assessment of the quality of studies included in systematic reviews of diagnostic accuracy is widely accepted and used. The selection process for this review resulted in at least half of included studies having used QUADAS. Overall, 85% (46 of 54) of reviews studied used a quality assessment, 61% (33) of reviews used QUADAS. Almost all reviews (42 of 46) used or adapted previously developed quality criteria, with only 4 studies developing their own criteria. Only 8 reviews did not use a formal quality assessment but 3 of these used quality items as inclusion criteria demonstrating awareness of the importance of study quality.

Use of QUADAS

Most reviews that used QUADAS assessed over 80% of QUADAS items. The item relating to patient spectrum (item 1) was the only item not omitted by any review, although four reviews modified this item. Modifications generally involved additional subcategories usually relating to study design. The most commonly omitted items were the availability of clinical information (item 12), avoidance of incorporation bias (item 7) and use of an appropriate reference standard (item 3). However, those reviews that omitted the item relating to reference standard generally restricted inclusion based on reference standard. Although some reviews added additional quality items there were no items that were consistently added across multiple reviews. QUADAS guidelines recommend adding review specific items as needed. Twelve reviews added their own quality items, although only 3 reviews added items that were review specific. 7 reviews added 13 items identified as potential additional quality items in Table 9.2 of the Cochrane DTA Handbook. Items added included: funding, prospective recruitment, proportion of patients recruited enrolled, inter-observer variability, evaluation of current test technology, reporting of definition of positive test result prior, determination of threshold prior to study commencement, administering of preventive intervention, whether the results were valid, description of setting for the test interpretation, patient or segment unit of analysis, and reporting of methods of analysis. One review not using QUADAS included a quality item that assessed whether a study evaluated diagnostic performance in a population other than the one used to derive the instrument (external validation).(19) This quality item is particularly relevant to reviews of diagnostic tests composed of multi-component scores such as patient questionnaires, where there are additional and large biases present in studies which report results from the same population that was used to derive or adapt multi-component scores.

The most commonly assessed item in reviews that did not use QUADAS was also patient spectrum. As with the QUADAS reviews, these reviews generally included additional subcategories relating to this item. Typical reviewers using QUADAS assessed quality using on average of twice as many quality items than those not using QUADAS. In reviews using QUADAS, a median of 14 quality items (IQR 13 to 15, range 5 to 19) were used in the quality assessment, of which a median of 1 (IQR 0 to 2, range 0 to 7) were non QUADAS items. In reviews not using QUADAS a median of 7 quality items (IQR 5 to 9) were used, of which a median of 5 (IQR 4 to 6) mapped to QUADAS items, and a median of 2 items (IQR 1 to 3) not mapping to QUADAS being added.

When QUADAS was not used for quality assessment, important aspects of quality were frequently omitted (Table 2). This was clearly demonstrated for item 6, where the key quality criteria "Did patients receive the same reference test regardless of the index test result?" was not used in any reviews using an alternative method of assessment to QUADAS.   Very few non-QUADAS reviews (two or less) assessed items relating to reporting of selection criteria (item 2), disease progression bias (item 4), incorporation bias (item 7), reference standard execution (item 9), availability of clinical information (item 12), and reporting of uninterpretable results/withdrawals (item 13).

(Table 2). It therefore appears that use of QUADAS has prompted a more complete assessment of study quality than reviews that have not used QUADAS.

Are QUADAS guidelines being followed or just the QUADAS checklist items?

Although reviews using QUADAS reference Whiting 2003, it was evident that many reviews were not following the QUADAS guidelines as published, although reviewers were using QUADAS checklist items. For example for item 1 of the QUADAS guidelines it is stated that reviewers should report the pre-specified criteria for an acceptable spectrum of patients with recruited patient characteristics. However almost none of the reviewers (6 of 6 reviews examined) reported item 1 criteria, although some reviews included reporting of individual study characteristics. Although this may be considered a poor reporting issue, as much as a misuse of QUADAS guidelines, the reader of a review is left with insufficient information to interpret the quality of studies when the definition of an acceptable spectrum is not provided. This is particularly evident when there is poor reporting of individual study characteristics and scores for individual QUADAS items for each study.

Reported problems with applying QUADAS items

Four reviews specifically reported difficulties in using QUADAS (Table 4). Problems scoring items 8 and 9 are to be expected, as assessment of sufficient technical detail may depend on the familiarity of the assessor with the test being used. Problems with items 4, 12 and 13, appear to be due to poor reporting in the primary diagnostic studies, although the time period in item 4 requires a subjective decision by reviewers.

Incorporation and reporting of quality

Although the proportion of reviews that used QUADAS to produce summary quality scores was lower than the proportion of non-QUADAS reviews reporting quality scores, this was still a significant proportion (around one third) given the explicit guidance within QUADAS not to use such scores. Studies that attempted to group studies based on quality (e.g. high and low) tended to do this based on summary quality scores rather than individual items considered to be of particular importance for their review. Methods used to report the results of the quality assessment was similar between QUADAS and non-QUADAS review and over time, when compared to the results of our previous review. Methods used to incorporate QUADAS into the review were also similar for

QUADAS and non-QUADAS reviews although it appeared that more recent reviews may tend to use regression analysis more than subgroup analysis to incorporate quality into the review compared to the reviews assessed in our original review.

3.5 Implications for QUADAS-2

- Consider modifying patient spectrum (item 1) –possible additional sub-categories e.g. study design, method of enrolment
- Possible item for omission or clarification – availability of clinical review bias (item 12) and incorporation bias (item 7)
- Possible items for inclusion: funding, prospective recruitment, proportion of patients recruited enrolled, inter-observer variability, evaluation of current test technology, reporting of definition of positive test result prior, determination of threshold prior to study commencement, administering of preventive intervention, whether the results were valid, description of setting for the test interpretation, patient or segment unit of analysis, and reporting of methods of analysis. In addition relevant to diagnostic tests composed of multi-component scores such as patient questionnaires, a possible item for inclusion is whether study results are from the same population used to derive or adapt a new multi-component score, or from an external population.
- Emphasise importance of following QUADAS guidance and not just using the checklist items
- Emphasis importance of avoiding use of summary scores
- Consider including explicit suggestions for overall rating of study quality and/or grouping studies based on quality

Chapter 4: Feedback from Reviewers

Key points

64 reviewers completed a questionnaire designed to gather feedback from reviewers who have used QUADAS

Positive features: coverage, ease of use, length/quick to complete, clarity, guidance documents, and the fact that it was evidence based.
Negative features: lack of consistency, need for modification to the review topic, problems with items 13/14 (uninterpretable results/withdrawals), poor reporting of primary studies, understanding and applying item 12 (availability of clinical information), lack of details for comparative studies, internal and external validity mixed up

Frequently omitted items: reporting of selection criteria (item 2), disease progression bias (11 reviewers), differential verification bias (6 reviewers), incorporation bias (item 7), execution of index test and reference standard (item 8 and 9).
*Ommissions: i*tems were rarely modified: no item modified by more than three reviewers.
Suggestions for additions: case-control design/split patient spectrum item, Items related to comparative studies, Observer variability/experience, Hypothesis (defined)

Despite explicit guidance not to produce summary scores, 20% of reviewers calculated these and third stratified findings based on quality.

General suggestions: expand QUADAS to handle *c*omparative tests, statistical correction for verification bias, and topics in which the reference standard consists of follow-up. Remove items related to the quality of reporting. Include some form of global rating of study quality, maintain the ability to modify QUADAS to address specific review questions, and extend QUADAS to prediction research.

QUADAS has been available for use in DTA reviews since 2003 and since then has been used in a large number of reviews covering a variety of topics. Although we have received some informal feedback we decided to develop a formal means of gathering feedback from reviewers who have used QUADAS to inform the development of QUADAS-2.

4.1 Objective

To gather structured feedback from reviewers who have used QUADAS

4.2 Methods

We developed a simple web-based questionnaire to gather structured feedback from reviewers who have used QUADAS. We invited all first authors of reviews indexed on DARE that have used QUADAS (96 reviews) and authors of published Cochrane DTA reviews and protocols to complete the questionnaire. In order to maximise response rates, we aimed to produce a questionnaire that was user-friendly, short and quick to complete.

4.3 Results

Full details of the questionnaire, including individual questions and detailed results, are presented in Appendix 4. We sent 118 e-mails inviting reviewers to complete the questionnaire and 64 respondents completed the questionnaire. The reviews covered a broad range of topics (full details in Appendix 4) including biochemical, histological, clinical and questionnaire tests.

General Details

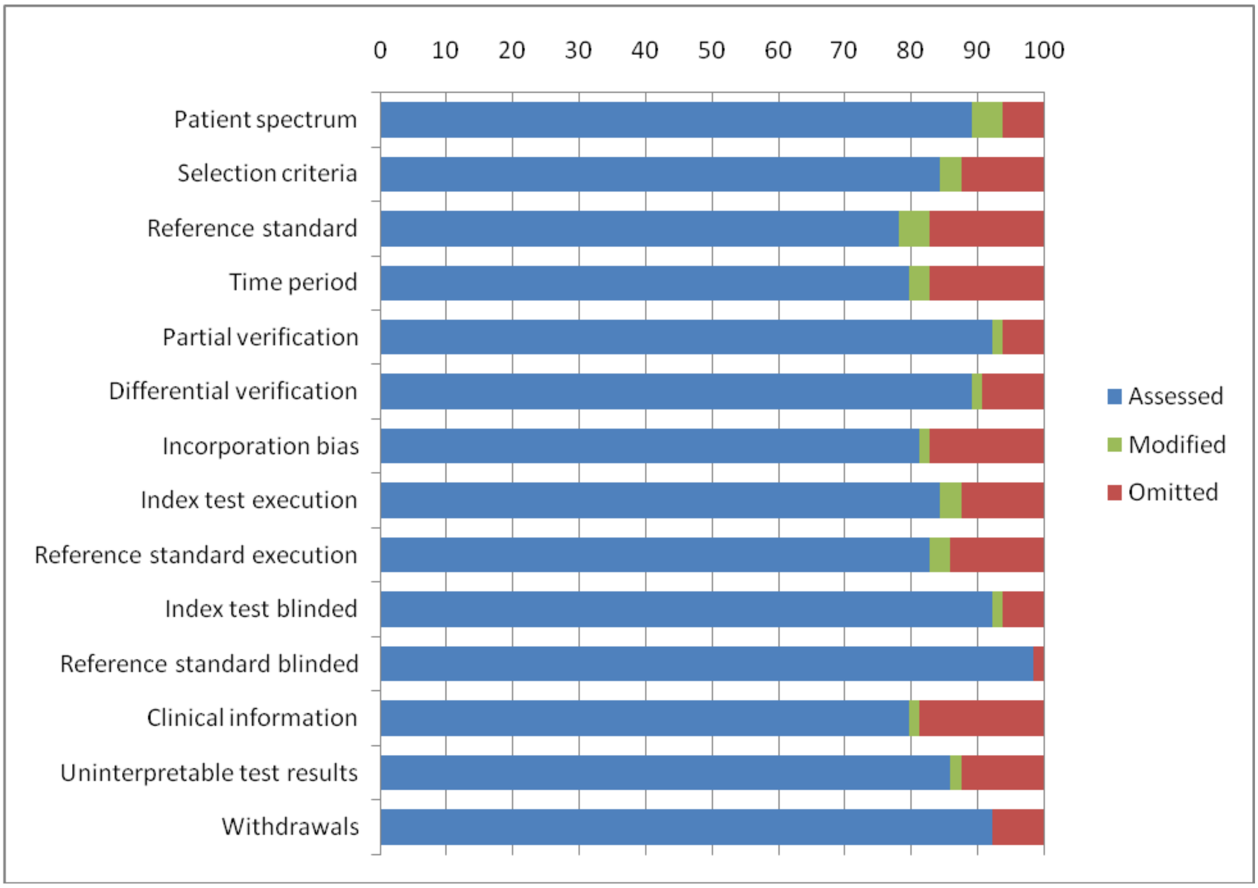
Most reviewers used QUADAS for non-Cochrane reviews (88%) and most (44%) only used QUADAS on one review, although 20% had used QUADAS in 4 to 5 reviews. Around 70% had previously conducted a quality assessment as part of a systematic review prior to using QUADAS, but for most of these (73%), this was a non-diagnostic review. There was substantial range in the amount of time that reviewers took to complete QUADAS: most reviewers took between 10 and 30 minutes but 5% took less than 5 minutes and another 5% took 1 to 2 hours. Almost all reviewers (89%) found the time taken to complete QUADAS acceptable, although 3 (5%) stated that they found the amount of time unacceptable and four were undecided. Of those that found the amount of time taken to complete QUADAS unacceptable, two took between 30 minutes and 1 hour and one took between 10 and 30 minutes. All of those who took between 1 and 2 hours to complete the assessment considered this to be an acceptable amount of time.

Use of QUADAS

Table 4.1 and Figure 4.1 summarise the number (%) of reviewers who assessed, omitted or modified each QUADAS item. Twenty seven reviews (42%) used QUADAS in its original format without any modifications or omissions. The number of reviewers omitting items ranged from one to 11 reviewers across items. Fewer reviewers modified items: the number of reviewers modifying a particular item ranged from 0 to 3 across QUADAS items. Reasons for modification or omission, where reported, are summarised below for each QUADAS item. Where reviewers modified questions by simply making them applicable to their reviews, we did not consider this to be a true modification and these reviewers were classed as having assessed this item for the purpose of analysis. On some occasions, QUADAS items were omitted because they were covered by the inclusion criteria (1 reviewer for patient spectrum, 8 for reference standard, and two for partial verification bias).

Table/Figure 4.1 Number (%) of reviewers who assessed, omitted or modified each QUADAS item.

Item		Assessed (%)	Omitted (%)	Modified (%)
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	57 (89)	4 (6)	3 (5)
2.	Were selection criteria clearly described?	54 (84)	8 (13)	2 (3)
3.	Is the reference standard likely to correctly classify the target condition?	50 (78)	11 (17)	3 (5)
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?	51 (80)	11 (17)	2 (3)
5.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	59 (92)	4 (6)	1 (2)
6.	Did patients receive the same reference standard regardless of the index test result?	57 (89)	6 (9)	1 (2)
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	52 (81)	11(17)	1 (2)
8.	Was the execution of the index test described in sufficient detail to permit replication of the test?	54 (84)	8 (13)	2 (3)
9.	Was the execution of the reference standard described in sufficient detail to permit its replication?	53 (83)	9 (14)	2 (3)
10.	Were the index test results interpreted without knowledge of the results of the reference standard?	59 (92)	4(6)	1 (2)
11.	Were the reference standard results interpreted without knowledge of the results of the index test?	63 (98)	1(2)	0
12.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	51 (80)	12 (19)	1 (2)
13.	Were uninterpretable/ intermediate test results reported?	55 (86)	8 (13)	1 (2)
14.	Were withdrawals from the study explained?	59 (92)	5 (8)	1 (2)



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

Omitted (4 reviewers)

One reviewer omitted this item as inclusion was restricted to studies that enrolled an appropriate patient spectrum, one stated that "a normal population was screened rather than a patient group", and one assessed external validity separately; the remaining reviewer did not report on the reason for omission.

Modified (3 reviewers)

Modifications included whether recruitment was consecutive (2 reviews) and the other stated that spectrum was also described in detail in a separate table.

Item 2: Were selection criteria clearly described?

Omitted (8 reviewers)

This item is not included in the Cochrane version of QUADAS and three reviewers cited this as the reason for omission, a further review stated that this item was excluded as it relates to reporting quality rather than methodological quality. Two reviewers stated that this item was not assessed (reasons not stated) but that data were extracted on selection criteria. The other two reviewer did not report reasons for omission.

Modified (2 reviewers)

Both reviews modified this item to consider the potential for bias rather than assessing reporting. Modified questions assessed were "Was inclusion of subjects based on the results of the index or comparator tests" and “Were inclusion/exclusion criteria applied consistently? Were consecutive eligible patients enrolled?”.

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

Omitted (11 reviewers)

Eight reviewers stated that specific reference standards were specified for inclusion and so this item was no longer relevant. One reviewer stated that there were multiple target conditions and another that there was no agreed reference standard. One reviewer did not report reasons for omission.

Modified (3 reviewers)

One reviewer stated that there was no agreed reference standard and so reference standard had to be considered as stated in the primary studies. The other stated that they were considering two outcomes in their review and so this item was included twice, once for each outcome. The third reviewer did not provide details of modifications.

Item 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?

Omitted (11 reviewers)

Five reviewers stated that this was omitted as both tests were performed at the same time. One reviewer stated that it was not relevant as they were assessing a genetic test, another that it was considered irrelevant in the context of their review, and another that the test was done in pregnant women with the reference standard assessed after birth. A further review stated that this item was incorporated into the item on reference standard by specificity that the reference standard had to be performed within 24 hours of the index test. Two reviewers did not report reasons for omission.

Modified (2 reviewers)

One reviewer adjusted this item to cover studies with follow-up as the reference standard and assessed the item “Was the follow-up appropriately long?”. The other review stated that index test was often performed on stored (blood) samples some time after reference standard (using the same blood, but before storage) and that this needed to be accommodated within this item.

Item 5: Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?

Omitted (4 reviewers)

Two reviewers stated that inclusion was restricted to studies in which all of the participants received the reference standard (i.e. those that avoided partial verification bias). One reviewer stated that all included studies only reported details on patients who received both the index test and reference standard. The fourth reviewer stated that different populations were used for validity.

Modified (1 reviewer)

One review stated that they separated the two different possibilities affecting partial verification bias: 1) random sample vs. non-random and 2) proportion of sample verified.

Item 6: Did patients receive the same reference standard regardless of the index test result?

Omitted (6 reviewers)

Three reviewers stated that inclusion was restricted to studies that used a single reference standard. One stated that there were difficulties in applying this when a genetic test is the reference standard and one stated that this is often unknown. One reviewer did not report on the reason for omission.

Modified (1 reviewer)

One reviewer stated that this item was split in two because it was possible that a different reference standard was applied but performance of the reference test was not related to the outcome of the index test.

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

Omitted (11 reviewers)

Three reviewers stated that the index test was always part of the reference standard and three stated that the index test could not be part of the reference standard. One reviewer stated that studies would have been excluded if incorporation bias were present. One reviewer stated that this item was considered irrelevant in the context and another stated that different populations were used. The other three reviewers did not report reasons for omission.

Modified (2 reviewer)

One reviewer stated that this item was often not applicable but not did not explain how the item was modified. The other reviewer stated that this item was only considered problematic in studies with short duration of follow-up when other clinical signs or symptoms may not have developed.

Item 8: Was the execution of the index test described in sufficient detail to permit replication of the test?

Omitted (8 reviewers)

One reviewer stated that this item was part of the inclusion criteria. One reviewer stated that all tests were commercial with package inserts or brochures describing the tests. One reviewer omitted this item as it related to the quality of reporting rather than methodological quality and another reviewer stated that they used the 11-item Cochrane version of QUADAS. One review stated that information was extracted on this but that it was not used as part of the quality assessment. The other three reviewers did not report reasons for omission.

Modified (2 reviewers)

One reviewer stated that they extended this item to assess whether the test was performed adequately according to international standards and the other stated that they were assessing two outcomes and so this item was assessed.

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

Omitted (9 reviewers)

One reviewer stated that this item was part of the inclusion criteria and so would have been scored as yes. One reviewer omitted this item as it related to the quality of reporting rather than methodological quality and another reviewer stated that they used the 11-item Cochrane version of QUADAS. One review stated that information was extracted on this but that it was not used as part of the quality assessment. One reviewer stated that they did not think that this item would help discriminate between good and less good studies. One reviewer stated that this was not usually an issue for their particular topic. The other three reviewers did not report reasons for omission.

Modified (2 reviewers)

One reviewer stated that they extended this item to assess whether the test was performed adequately according to international standards and the other stated that they were assessing two outcomes and so this item was assessed.

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

Omitted (4 reviewers)

Two reviewers stated that the index test would always be performed before the index test, one stated that the index test was objective, and another stated that different populations were used.

Modified (1 reviewer)

One review stated that they also evaluated whether the evaluation of the index text was blinded to the results of the comparator test and *vice versa.*

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

Omitted (1 reviewer)

One reviewer stated that different populations were used.

Modified

None of the reviewers modified this item.

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

Omitted (12 reviewer)

Four reviewers stated that this item was not considered relevant in the context of their review. One reviewer stated that information was always present in the studies included in their review, one review stated that they did not understand this item, and another stated that there was no way to get this information from the included studies. One reviewer stated that they asked if there was blinding to clinical data, to emphasize internal validity over external validity. Four studies did not report on reasons for omission.

Modified (1 reviewer)

Details of the modification were not reported.

Item 13: Were uninterpretable / intermediate test results reported?

Omitted (8 reviewers)

Two reviewers stated that the design of the studies meant that there were no intermediate results. Three reviewers stated that this item was not applicable. One reviewer replaced this item (and item 14) with "Were at least 85% of patients accounted for?". The other two reviewers did not report on reasons for omission.

Modified (1 reviewer)

One reviewer stated that this item needs more details on how this can be more scored more precisely given the possible bias if indeterminate results are removed or classed as positive or negative.

Item 14: Were withdrawals from the study explained?

Omitted (5 reviewers)

One reviewer stated that details of missing values were included in the data extraction table but were not scored as a QUADAS item and another stated that withdrawals were not mentioned in the studies and that they only present the patients who received both tests. One reviewer replaced this item (item 13) with “Were at least 85% of patients accounted for?”. The other two reviewers did not report on reasons for omission.

Modified (1 reviewer)

One reviewer modified this item to “Were withdrawals from the study documented at all?”.

Inter-rater reliability

Ten reviewers stated that they assessed inter-rater reliability. However, three of these did not provide any quantification of the level of agreement and one reviewer stated that quality assessment is ongoing and so inter-rater reliability has not yet been quantified. Absolute agreement was reported by three reviewers and ranged from 50% to absolute agreement. Kappa statistics were reported by three reviewers and ranged from 0.53 to >0.75. One of these reviewers stated that minimal conferencing yielded near perfect agreement.

Guidance and training

Most reviewers (89%) stated that they had read the QUADAS background document or the relevant Cochrane handbook chapter (27%). Of those that did not read the QUADAS background document, two stated that they were unaware of its existence but one of these reported having read the relevant Cochrane handbook chapter. Five reviewers stated that they did not read the background document but were aware of its existence and two of these stated that they read the relevant Cochrane handbook chapter. Thus all but four of the reviewers read one of the guidance documents on QUADAS. Although the majority of respondents found the background document easy to understand (87%), seven highlighted some problems with it. One review found the definitions of differential and partial verification difficult to understand and another had problems with the some explanations of the items relating to selection criteria and reference standard. One reviewer said they found it generally easy to understand but suggested that additional examples may have been helpful. One reviewer said that it was generally easy to understand but not when assessing a genetic test. One reviewer reported that it was “somewhat” easy to understand but that it took a very long time for research assistants to grasp.   Another review mentioned some issues specific to scoring items 2 and 12 for their review. The final reviewer stated that it remains vague how to score items and that the document is open to a lot of interpretation.

Twenty percent of reviewers sated that they did not use any guidelines when scoring QUADAS and a further 28% only referred reviewers to existing guidance documents. Around 30% of reviewers stated that they adapted existing guidance documents to make them specific to their reviews and 22% produced their own scoring guidelines.

The majority of reviewers (66%) had not received any formal training in QUADAS, although most (69%) stated that this would be helpful.   Nine reviewers had attended a workshop on QUADAS at a Cochrane Colloquium, three had attended training aimed at Cochrane Review Groups, two had received a workshop training session in Amsterdam, and one had attended a workshop on quality assessment at a symposium. Seven reviewers stated that they had received other training, this included hands on training by Cochrane expert (2 reviewers), attendance at symposia/conferences on diagnostic accuracy studies, reading, lecture on QUADAS as part of an MSc course, training by expert within the reviewers own institution and “various”. Twenty seven reviewers stated that internal training sessions were organised to ensure that reviewers applied the tool consistently. These sessions tended to include agreement of quality criteria, piloting the quality assessment, discussion of discrepancies after pilot quality assessment, practice with relevant studies followed by discussion.

Incorporating QUADAS results into the review

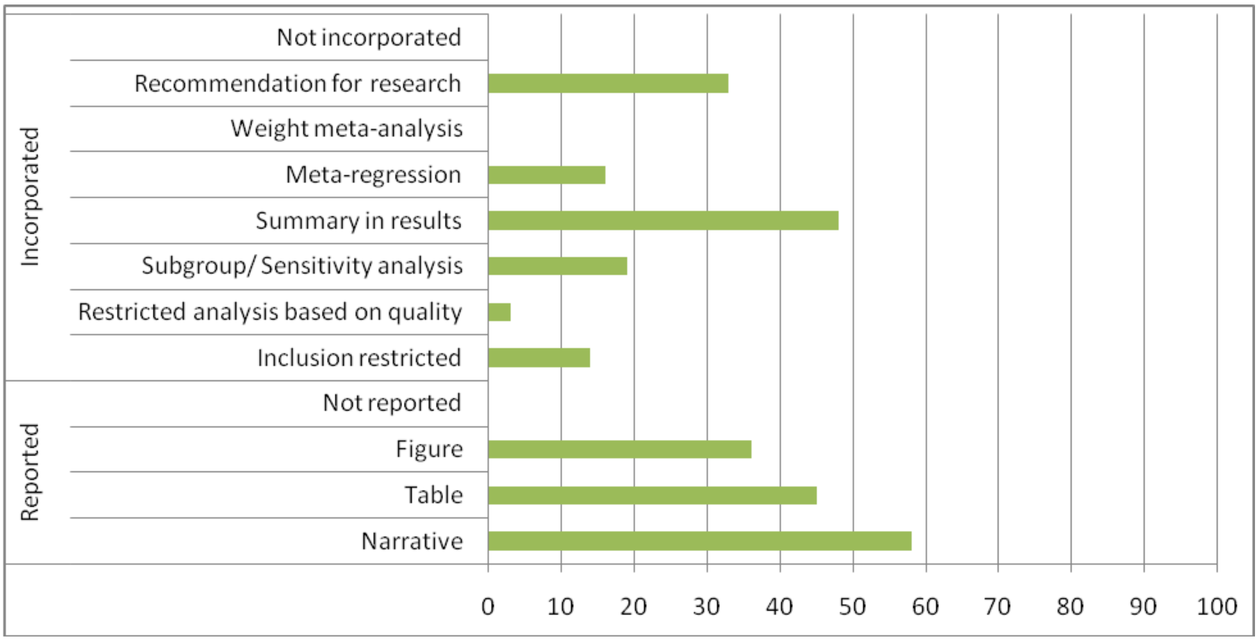
Despite clear guidance in the background documents accompanying QUADAS against calculating summary scores, 20% of reviewers reported using QUADAS to calculate a summary quality score. Most studies summed “yes” rating to get a summary score, but some used more complicated variations scoring “yes” as 1 or 2, “no” as -1 or 0 and “unclear” as 1, 0 or -0.5. None of the studies assigned different weight to the individual QUADAS items when calculating the summary score.

Nineteen of the reviewers used QUADAS to stratify studies according to quality. Ten reviewers stratified studies into different grading of quality based on the summary quality scores. Thresholds used to define a "high quality study" varied substantially between studies ranging from 8 to >12. Three reviewers stated that they performed subgroup analysis based on specific QUADAS items. Three reviewers based the stratification on QUADAS items which they considered to be most important for their reviews.

Methods used to report the results of the QUADAS assessment and to incorporate this into the review are summarised in Table/Figure 4.2. Although 13 reviewers stated that they used other methods, details of methods reported fitted into the categories outlined in Table/Figure 4.2.

Table /Figure 4.2: Details on how quality was reported and incorporated in the review

Method of reporting/incorporating study quality		**Number of reviews (%)**
How were the results of the QA reported?	Narrative	37 (58)
	Table	29 (45)
	Figure	23 (36)
How were the results of the QA incorporated?	Inclusion in review	9 (14)
	Inclusion in primary analysis	2 (3)
	Subgroup/ Sensitivity analyses	14 (22)
	Summary in results	31 (48)
	Meta-regression	10 (16)
	Weight meta-analysis	0
	Recommendation for research	21 (33)
	Other	13 (20)



Rating of QUADAS

Reviewers were asked to rate QUADAS on a 5-point scale ranging from very poor to very good for whether they felt that QUADAS included all important items, ease of use, clarity of instructions and validity. The number of reviewers assigning each rating to each of these items is summarised in Table 4.3. Most reviewers (70% to 89%) reviewers rated QUADAS as good or very good on each of these items. The ratings for including all important items and ease of use were very good with no reviewers rating these items as poor or very poor. Two reviewers rated the clarity of instruction as poor. Both of these had used the Cochrane background document, one rated this as easy to understand and the other as “somewhat easy to understand”. One reviewer rated the overall validity of QUADAS, its ability to help differentiating between studies of different qualities, as very poor and three reviewers rated this as poor. However, the reviewer that rated this item as “very poor” stated that they did so because this question had to be answered and they stated that they have no way of knowing whether QUADAS can make this distinction.

Table 4.3 Number (%) of reviewers who assigned rating ranging from very poor to very good for features relating to QUADAS

Feature	**Very Poor**	**Poor**	**Average**	**Good**	**Very Good**
Inclusion of all important items	0	0	7 (11)	32 (50)	24 (39)
Ease of use	0	0	16 (25)	34 (53)	14 (22)
Clarity of instructions	0	2(3)	15 (23)	31 (48)	16 (25)
Validity	1 (2)	3 (5)	15 (23)	30 (47)	15 (23)

Aspect of QUADAS that reviewers liked

Reviewers highlighted a broad range of features that they liked about QUADAS. The most commonly reported were coverage (19 reviewers), ease of use (11 reviewers), length/quick to complete (7 reviewers), clarity (5 reviewers), guidance documents (4 reviewers), and the fact that it was evidence based (2 reviewers). Items highlighted by single reviewers included coverage of external validity, "reliably subjective", acknowledges need for modification, the rating of yes/no/unclear, "good starting point", prompted interesting discussion, forces authors to assess sample characteristics, and "it exists".

Aspect of QUADAS that reviewers do not like

There was substantial variation in aspects of QUADAS that reviewers did not like with few features picked up as problematic by more than one reviewer. Issues that were raised by multiple reviewers were subjectivity in interpretation/lack of consistency between raters (7 reviewers), the need for modification to the reviewer topic (4 reviewers), problems with items 13/14 (uninterpretable results/withdrawals)(3 reviewers), poor reporting of primary studies (3 reviewers), understanding and applying item 12 (availability of clinical information)(3 reviewers), lack of details of comparative studies (2 reviewers), internal and external validity mixed up (2 reviewers), some items are often scored unclear (2 reviewers), difficult to always rate yes/no/unclear / need to for additional item of "not applicable" (2 reviewers). Other issues raised were that it is difficult to use without methodological expertise, can be difficult to understand, some items are "reporting items", lack of a question relating to case-control designs, and missing details on sample size.

Suggestions for improving QUADAS

A broad variety of helpful suggestions were made for improving QUADAS. We have grouped these into suggestions relating to quality items, to guidance and to general features of the tool:

Items

The following items were suggested for inclusion in QUADAS:

- Use of case-control design

- Were withdrawals explained
- Items related to comparative studies (3 reviewers)
- Observer variability/experience (2 reviewers)
- Prospective/retrospective data collection
- Hypothesis (defined) (2 reviewers)
- Unbiased patients selection
- Adequate statistical methods
- Sample size
- Reporting of data on existing tests
- Conflicts of interest
- Split spectrum into 2 items (exact items not specified)
- Technological status of index test

General features

Recommendations for general features of QUADAS included modifications so that it could handle the following situations:

- Comparative tests
- Statistical correction for verification bias
- Topics in which the reference standard consists of follow-up,
- Remove items related to the quality of reporting

Suggestions also included having some form of global rating of study quality, to maintain the ability to modify QUADAS to address specific review questions, and to extend QUADAS to prediction research.

Guidance

Two reviewers suggested that it would be helpful to include more examples in the scoring guidance for QUADAS, one reviewer expressed a specific desire to include examples related to laboratory tests. A suggestion was to have some way of gathering together the different modifications to QUADAS that reviewers have made for their reviews, possibly via an online database. One reviewer stated that it would be helpful to have guidance on the likely direction of the different sources of bias. Another reviewer stated the need to emphasise that QUADAS should be adapted specifically for individual reviews. One reviewer requested guidelines on how to produce summary quality scores.

Final comments

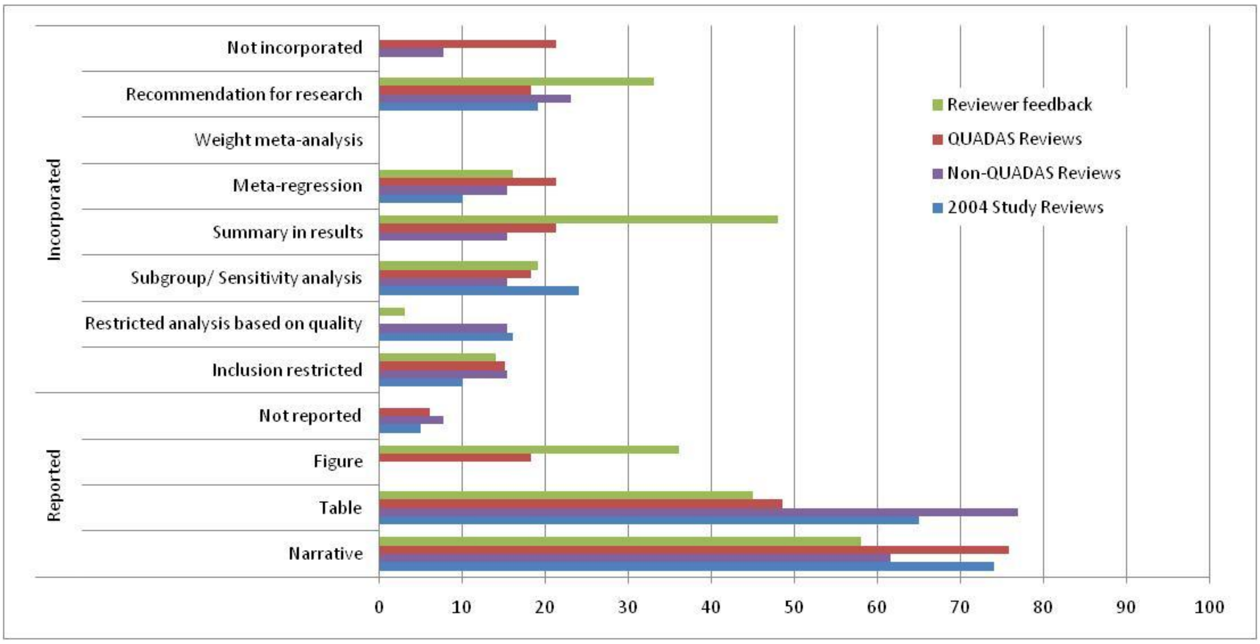
All reviewers stated that they would use QUADAS again. Final comments were generally complementary about QUADAS and the work to update it.

4.4 Summary

Feedback from reviewers was generally positive with all reviewers stating that they would use QUADAS again. The majority of reviewers (70% to 89%) rated QUADAS as good or very good for coverage, ease of use, clarity and validity. Most reviewers found the length of time taken to complete QUADAS acceptable. When reviewers were asked to indicate features associated with QUADAS that they liked the following were highlighted by multiple reviewers: coverage, ease of use, length/quick to complete, clarity, guidance documents, and the fact that it was evidence based.

Reviewers were also asked to indicate features that they did not like. There was less consistency in ratings for this with the following items highlighted by more than one reviewer: subjectivity in interpretation/lack of consistency between raters, the need for modification to the review topic, problems with items 13/14 (uninterpretable results/withdrawals), poor reporting of primary studies, understanding and applying item 12 (availability of clinical information), lack of details of comparative studies, internal and external validity mixed up, some items are often scored unclear, difficult to always rate yes/no/unclear / need to for additional item of “not applicable”. Items omitted by more than five reviewers were reporting of selection criteria (item 2), disease progression bias (11 reviewers), differential verification bias (6 reviewers), incorporation bias (item 7), execution of index test and reference standard (item 8 and 9). Although use of an appropriate reference standard was also frequently omitted, most reviews that did so restricted inclusion based on reference standard. Items were rarely modified with no item modified by more than three reviewers. Despite three reviewers highlighting items 13 and 14 (uninterpretable results and withdrawals) as problematic, these were rarely omitted or modified. Although most reviewers were aware of the existence of guidance relating to QUADAS, 20% did not use specific guidance, either the existing background documents or guidance developed specifically for their reviews, when using QUADAS. Despite explicit guidance accompanying QUADAS not to produce summary scores, 20% of reviewers calculated these. Around a third of reviewers stratified findings based on quality and over half of these used summary scores rather than individual item(s) to do so. When reviewers were asked to suggest improvements to QUADAS a number of additional items, features and improvements to guidance were proposed. Reviewers reported similar methods of incorporating the results of their QUADAS assessment into the review as we found in our reviews of the published literature (Chapter 3). Figure 4.2 summarises how quality was incorporated into the results of reviews based on the reviews included in Chapter 3 and the questionnaires evaluated in this chapter.

Figure 4.3: Details on how quality was reported and incorporated in the review based on reviews in Chapter 3 and reviewers' questionnaire responses



4.5 Implications for QUADAS-2

- Consider modifying patient spectrum by adding the following sub-categories: use of case-control design, prospective/retrospective data collection, unbiased patients selection
- Possible items for inclusion: Observer variability/experience, Hypothesis (defined), Adequate statistical methods, Sample size, Reporting of data on existing tests, Conflicts of interest, Technological status of index test

- Possible items for omission or clarification: availability of clinical information (item 12), incorporation bias (item 7), reporting of uninterpretable results and/or withdrawals (items 13 and 14)
- Consider expanding QUADAS to cover the following situations: comparative tests, statistical correction for verification bias, topics in which the reference standard consists of follow-up, remove items related to the quality of reporting
- Emphasise importance of avoiding use of summary scores
- Emphasise importance of developing review specific scoring guidance
- Consider including explicit suggestions for overall rating of study quality and/or grouping studies based on quality
- Consider including additional examples in the scoring guidance covering a broader variety of topics
- Consider providing an online learning resources that is continually updated based on reviewers' experience of using QUADAS

Chapter 5: Sources of Variation and Bias in Studies of Diagnostic Accuracy: an updated systematic review

Key points

The original review included 55 studies; we included an additional 46 studies giving a total of 101.

There was considerable evidence for the effects of demographic features, distorted selection of participants, disease prevalence, disease severity, inappropriate reference standard, partial verification bias, and observer variation.

There was adequate evidence for the effects of differential verification bias, review bias, and clinical review bias.

There was some evidence for the effects of prior testing, test technology, test execution, disease progression bias, incorporation bias, instrument variation, withdrawals, arbitrary choice of threshold and sample size.

There was no evidence to support the effects of inappropriate handling of uninterpretable test results or treatment paradox on estimates of test performance.

In 2004 we published a systematic review on sources of bias and variation in studies of diagnostic tests.(5) The goal of this study was to classify the different sources of variation and bias, describe their effects on test results, and provide a summary of the available evidence of the effects of each source of bias and variation.

The original review included 55 studies published from 1963 to 2000. Nine studies were systematic reviews, 16 studies used an experimental design, 22 studies were diagnostic accuracy studies, and 8 studies used modelling to investigate the theoretical effects of bias or variation. The studies were concentrated in seven areas of bias and variation: demographic features (10 studies), disease prevalence (6 studies), disease severity (6 studies), inappropriate reference standard (8 studies), partial verification bias (20 studies), clinical review bias (9 studies), and observer variation (8 studies). The best-documented effects of bias and variation were found for demographic features, disease prevalence and severity, partial verification bias, clinical review bias, and observer and instrument variation. For other sources, such as distorted selection of participants, absent or inappropriate reference standard, differential verification bias, and review bias, the amount of evidence was limited. Other sources of bias commonly believed to affect studies of diagnostic test performance, such as incorporation bias, treatment paradox, arbitrary choice of threshold value, and dropouts, were not considered in any studies.

5.1 Objectives

To update the original review to provide an up to date summary of the evidence of the effects of sources of bias and variation on estimates of diagnostic accuracy.

5.2 Methods

Literature searches

The searches for the original review were carried out from database inception to 2001; we updated these searches. We searched MEDLINE, EMBASE, BIOSIS, the Cochrane Methodology and DARE from 2001 to April 2010. Full details of the search strategy are provided in Appendix 5.1. Search terms included *sensitivit**, *mass-screening*, *diagnostic-test*, *laboratory-diagnosis*, *false positive**, *false negative**, *specificit**, *screening*, *accuracy*, *predictive value**, *reference value**, *likelihood ratio*', *sroc*, and *receiver operat* characteristic**. We carried out a citation search to identify studies that cited key papers (Begg (1987)(20), Lijmer (1999)(18) and Whiting(2004)(5)). The results of the searches were screened independently by two reviewers.

Inclusion Criteria

We adopted the same inclusion criteria as used in the original review. All studies with the main objective of addressing bias or variation in the results of diagnostic accuracy studies were eligible for inclusion. Studies of any design, including reviews, experimental studies and theoretical modelling, and any topic area were eligible. Studies had to investigate the effects of bias or variation on measures of test performance, such as sensitivity, specificity, predictive values, likelihood ratios, and diagnostic odds ratios, and indicate how a particular feature may distort these measures. Inclusion was assessed by one reviewer and checked by a second; discrepancies were resolved through discussion or referral to a third reviewer where necessary.

Data Extraction

One reviewer extracted data on the following parameters: study design, objective, sources of bias, variation or applicability investigated, and the results for each source. A second reviewer checked the data extraction. Discrepancies were resolved by consensus or consultation with a third reviewer.

Classification of sources of bias and variation

Our original review classified each item as a possible source of "bias" or "variation". A design feature may bias the results of a study if it leads to a systematic departure from the "true" result. A source of bias has the potential to produce inaccurate and misleading results. In contrast a source of variation is a feature that can result in differences in estimates of accuracy across studies but does not bias the results of a study. For example, differences in test protocol or differences in study populations can produce different estimates of accuracy. These are not biased estimates but the results may only be applicable to the particular test protocol or population in which the study was conducted. We adopted the classification of items as sources of bias or variation used in our original review (Table 5.1).

Data Synthesis

We divided the different sources of bias, variation and applicability into the groups shown in Table 5.1, which provides a brief description of each source of bias and variation; more detailed descriptions are available elsewhere.(5) Results were stratified according to the source of bias or variation. Studies were grouped according to study design. We classified studies that used actual data from one or more clinical studies to demonstrate the effect of a particular study feature as experimental studies, diagnostic accuracy studies, or systematic reviews. Experimental studies are those designed specifically to test a hypothesis about the effect of a certain feature, for example, rereading sets of radiographs while controlling (manipulating) the overall prevalence of abnormalities. Studies that used models to simulate how certain types of biases may affect estimates of diagnostic test performance were classified as modelling studies. These studies were considered to provide theoretical evidence of bias or variation.

5.3 Results

The literature searches identified a total of 4783 references. Of these, 123 studies were considered potentially relevant and were assessed for inclusion, and 46 met inclusion criteria. A further 55 studies were included in our original review and are also included in this update review. Thus a total of 101 studies were included. The year of publication of the included studies ranged from 1963 to 2010. Individual study results are presented in Appendix 5.1. A narrative analysis was provided in five studies and a statistical analysis in the remaining 96 studies. Ninety one studies provided empirical evidence of bias and fifteen provided theoretical evidence (five studies provided both forms of evidence). A diagnostic accuracy design was used in 39 studies, of which 22 were prospective and 17 retrospective. Twenty two studies were systematic reviews (three meta-reviews) and 17 studies used an experimental design.

Spectrum composition

Variation by clinical and demographic subgroups

Twenty six studies investigated the effects of variations in clinical and demographic features on test performance, 16 diagnostic accuracy studies, 2 modelling studies, and eight reviews.   Nine studies found no evidence of an association between the features investigated and estimates of accuracy. All other studies provided empirical evidence of an association. A variety of possible sources of variation were investigated including gender, age, weight, history of prior disease, disease related features, smoking, co-morbidities, race/ethnicity, medication use, symptoms, BMI, menopausal status, and educational level.   The direction of the association varied between studies with sensitivity more commonly affected than specificity. Fourteen studies reported an association of the factors investigated and sensitivity, eight studies reported associations with specificity (7 also reported an association with sensitivity), and three studies reported an association with overall accuracy.

Distorted selection of participants

Sixteen studies looked at the effects of distorted selection of participants on test performance, two diagnostic accuracy studies, one modelling study, and 13 reviews (3 meta-reviews). A variety of different features related to patient selection were considered, with some studies assessing multiple features:

Study design (case-control versus cohort) (7 studies): Three studies reported increased estimates of overall accuracy in case-control studies compared to cohort studies, one of these also reported increased estimates of sensitivity and specificity. One study reported greater sensitivity but no effect on specificity, and two found no association with estimates of accuracy. A further study provided theoretical evidence that there was no difference between estimates of accuracy derived from nested case-control samples drawn from a single cohort compared to estimates of accuracy for the whole cohort.

Prospective data collection (4 studies): Two studies reported that retrospective studies increased accuracy compared to prospective studies, and two found no association with accuracy.

Consecutive patient enrolment (2 studies): Two studies compared estimates of accuracy from consecutive samples to those from non-consecutive samples and found no association with accuracy.

Other features related to recruitment (6 studies): Two studies found no association between accuracy and avoidance of a limited challenge group, one study found that failure to describe patient spectrum resulted in increased accuracy, one study reported that selection based on referral for index test decreased accuracy, one study found that in vivo studies increased accuracy compared to in vitro studies, and one study found that appropriate patient selection lead to increased sensitivity and specificity.

Disease prevalence

Fifteen studies looked at the effect of disease prevalence, eight diagnostic accuracy studies, one experimental study, one modelling study, and five reviews. All but one of the studies found associations between accuracy and disease prevalence. Four studies found that sensitivity increased and specificity decreased with increasing disease prevalence, one found that both sensitivity and specificity increased, two found increased sensitivity but no effect on specificity, one reported decreased sensitivity but did not assess the effect on specificity, one reported decreased specificity but did not assess sensitivity, two reported an association with overall accuracy, one review found that increasing prevalence increases the positive predictive value and decreases the negative predictive value, and a review reported that the direction and magnitude of the effect varied across studies. The final study reported that when prevalence is low, overall accuracy more closely resembles specificity; when prevalence is high, overall accuracy more closely resembles sensitivity.

Disease severity

Thirteen studies looked at the effect of disease severity, 7 diagnostic accuracy studies, 1 modelling study and five reviews. Eleven studies reported increased sensitivity and either did not assess the effect on specificity or found no association with specificity, on reported that disease severity was associated with accuracy, and one study found no association between disease prevalence and accuracy.

Prior testing

Three diagnostic accuracy studies assessed the influence of prior testing on estimates of accuracy. Two studies found no effect and the other reported increased sensitivity and decreased specificity.

Test protocol: materials and methods of testing

Change in technology of index test

Eight studies, two diagnostic accuracy studies and six reviews, looked at the effects of a change in the technology of the index test on test performance. Four studies found no association between test technology and test performance. Three studies found that improvements in test technology (automation, greater bronchial lavage volume, and higher transducer performance) resulted in increased sensitivity; one study also reported increased specificity, one reported decreased specificity and the other did not assess the effect on specificity. The final study, a review, found that accuracy was high in studies that used specific MRI imaging techniques.

Test execution

Four studies looked at the effects of execution of tests. Two studies found no association with different methods of test execution and accuracy estimates and one found no association between reporting of test execution and accuracy. One review found that failure to describe the index and reference standard execution biases estimation of test performance and provided empirical evidence of bias.

Disease progression bias

Four reviews assessed the effects of disease progression bias on test performance. Three found no effect on estimates of accuracy; one found that delayed verification resulted in decreased accuracy.

Treatment paradox

One meta-review assessed the effects of treatment paradox but found no association with overall accuracy.

Selection and execution

Absent or inappropriate reference standard

Eighteen studies looked at reference standard error bias, 2 meta-reviews, 10 reviews, 4 modelling studies, and 2 diagnostic accuracy studies. Ten studies found empirical evidence of bias, four found theoretical evidence and four found no association between estimates of accuracy and reference standard. The direction of the association varied between studies. Eight of the ten empirical studies found an association with sensitivity, two of these also found an association with specificity. The other two studies reported an association with overall accuracy. One study provided theoretical evidence suggesting that with imperfect reference standards specificity is most accurately estimated at low disease prevalence and sensitivity at high disease prevalence, and that considerable errors in estimates exist, even when the reference standards has close to perfect performance. The second theoretical study found that inaccurate reference standards lead to underestimation of test performance when the diagnostic test errors are statistically independent and overestimation when they are dependent. The other two theoretical studies found that test performance is underestimated when the test being evaluated is more accurate than the reference standard.

Partial verification bias

Thirty studies investigated the effects of partial verification bias, 12 diagnostic accuracy studies, 6 modelling studies, and 12 reviews (including 3 meta-reviews). Six studies found no evidence of bias on either overall accuracy (n=4), sensitivity and specificity (n=1) or any of these outcomes (n=1); five provided theoretical evidence of bias, one provided both theoretical and empirical evidence of bias, and the remaining 18 studies provided empirical evidence of bias. The effects of verification bias differed between studies although 12 studies reported increased sensitivity and decreased specificity. A further 3 studies reported increased sensitivity, one of these also found increased specificity and two found no association with specificity. Three studies reported decreased sensitivity, one of these also reported decreased specificity, one reported increased specificity and one found no association with specificity. Three studies found an association with specificity (increased in one, decreased in two) but no association with specificity. Two studies reported that overall accuracy was increased in the presence of verification bias and one found an association with overall accuracy.

Differential verification bias

Eight studies looked at differential verification bias, one diagnostic accuracy study and seven reviews (2 meta-reviews). Three found no association with accuracy. Two reviews reported an association with sensitivity (increased in one, direction of association not reported in the other), one of these also reported an increase in specificity. Two reviews reported that overall accuracy was increased in the presence of verification bias, and one review reported that there was a “potential for bias”.

Interpretation

Review bias (test and diagnostic)

Twelve reviews, including three meta-reviews, assessed review bias. Five studies assessed test review bias, three of these found no evidence of bias, one found that sensitivity and overall accuracy were increased and one found that sensitivity was decreased. Four studies assessed diagnostic review bias, one reported no association with accuracy, one reported increased sensitivity and one reported increased overall accuracy. Two studies assessed the effect of “double blinding”, both reported no association with overall accuracy. Five studies did not specify the type of review bias considered, two of these found no association of blinding with accuracy, two reported increased sensitivity and one of these also reported increased specificity, and one reported increased overall accuracy.

Clinical review bias

Thirteen studies looked at the effects of clinical review bias. Four studies found no difference in test performance between those tests interpreted with and without clinical history. Six studies found that sensitivity was improved when test results were interpreted with clinical history, two of these reported that specificity was decreased, the other did not assess specificity or found no effect. Three studies did not assess the effects on sensitivity or specificity but reported greater overall accuracy when tests were interpreted with clinical information; one of these was a modelling study and provided theoretical evidence of bias.

Observer variation

Fourteen studies looked at observer variation, two diagnostic accuracy studies, eight experimental studies and four reviews. Both diagnostic accuracy studies reported empirical evidence of bias. One found that sensitivity/overall accuracy was greater when for experts compared to non-experts. Seven of the eight experimental studies provided empirical evidence of bias for inter-observer variability and two also found evidence of intra-observer variability; one of these reported that inter-observer variability was greater than intra-observer variability. Two studies found that more experienced reviewers, or experts, provided greater sensitivity, while another found that experience was not related to inter-observer variability. Two of the reviews found evidence of bias, the other two found no association between observer experience and sensitivity/specificity. One of the reviews found that the observers’ threshold for interpreting a positive EEG was associated with accuracy, the other found greater accuracy when scans were interpreted by experts and multiple observers.

Instrument variation

Two reviews assessed instrument variation. One found that overall accuracy decreased when diagnosis was made based on experimental studies that involved presentation of slides compared to when the diagnosis was made to face. The other found no difference in accuracy between different laboratory methods.

Incorporation bias

Two reviews (one meta-review) assessed incorporation bias. The meta-review found no association with accuracy, the review found that sensitivity increased and specificity decreased in the presence of incorporation bias.

Analysis

Precision (sample size, variation by chance)

Two reviews assessed the influence of sample size on accuracy. One found no association and one found increased accuracy in studies with <30 patients.

Inappropriate handling of uninterpretable/indeterminate/intermediate test results

Three studies, two reviews and one diagnostic accuracy study, looked at the effects of uninterpretable test results. One study stated that a large proportion of results would be excluded if unsatisfactory test results were excluded, but provided no evidence as to how this may lead to biased estimates of test performance. One review found no association between treatment of equivocal or non-diagnostic tests and overall accuracy and the other found no association with sensitivity.

Dropouts

One review concluded that studies reporting on the number of excluded patients and drop-outs had lower sensitivity than those that did not.

Post hoc choice of threshold value

Five studies assessed the influence of threshold. Two reviews (one meta-review) found no association between method of threshold selection and accuracy, one diagnostic accuracy reported increased sensitivity when threshold was selected based on a fixed value of specificity, and two modelling studies provided theoretical evidence that data-driven threshold selection increases sensitivity and specificity compared to using a pre-specified threshold and that the size of the bias is greater with smaller sample sizes.

Table 5.1: Definition of sources of bias and variation with number of studies providing empirical, theoretical or no evidence of bias for each source of bias (numbers in brackets indicate number of studies providing each type of evidence from the original bias review)

Category of bias	**Source of bias or variation**	**Description**	**Evidence of effect of bias (number of studies)**		
			Emp-irical	**Theor-etical**	**No evidence**
Spectrum composition	Demographic features (variation)	Tests may perform differently in different populations. Demographic features may therefore lead to variations in estimates of test performance.	17 (9)	0	9 (1)
	Distorted selection of participants (variation)	The selection process determines the composition of the study population. If the selection process does not aim to include a patient spectrum similar to the population in which the test will be used in practice then the results of the study may have limited applicability	11 (3)	0	5 (1)
	Disease prevalence (variation)	The prevalence of the target condition varies according to setting and may affect estimates of test performance. Context bias, the tendency of interpreters to consider test results more frequently abnormal in settings with higher disease prevalence, may also impact estimates of test performance.	13 (6)	0	1 (0)
	Disease severity (variation)	Differences in disease severity between studies may lead to differences in estimates of test performance.	12 (6)	0	1 (0)
	Prior testing (variation)	Different in prior test results may lead to differences in estimates of test accuracy.	1 (0)	0	2 (0)
Test protocol: material and methods	Test technology (variation)	When the characteristics of a diagnostic test change over time, owing to technological improvement or to the experience of the operator of the test, estimates of test performance may be affected.	4(1)	0	4(1)
	Test execution (variation)	A sufficient description of the execution of index and reference standards is important because variation in measures of diagnostic accuracy can be the result of differences in test execution	2(1)	0	3(1)
	Disease progression bias (bias)	Disease progression bias occurs when the index test is performed an abnormally long time before the reference standard, so the disease is at a more advanced stage when the reference standard is performed	1(0)	0	3(1)
	Treatment paradox (bias)	Treatment paradox occurs when treatment is started on the basis of the knowledge of the results of the index test, and the reference standard is applied after treatment has started	0	0	1(0)
Reference standard and verification procedure	Inappropriate reference standard	When errors of imperfect reference standard(s) bias the measurement of diagnostic accuracy of the index test.	10(4)	4(4)	4(0)
	Differential verification bias (bias)	When part of the index test results are verified by a different reference standard.	5(2)	0	3(0)

Category of bias	**Source of bias or variation**	**Description**	**Evidence of effect of bias (number of studies)**		
			Empirical	**Theoretical**	**No evidence**
	Partial verification bias (bias)	When only a selected sample of patients that underwent the index test is verified by the reference standard.	18(15)	6(3)	6(3)
Interpretation (reading process)	Review bias (bias)	When interpretation of the index test or reference standard is influenced by knowledge of the results of the other test. Diagnostic review bias occurs when the results of the index test are known while interpreting the reference standard. Test review bias occurs when results of the reference standard are known while interpreting the index test.	7(3)	0	5(1)
	Clinical review bias (bias)	The availability of information on clinical data, such as age, sex and symptoms, during interpretation of test results may affect estimates of test performance.	8(8)	1(0)	4(1)
	Incorporation bias (bias)	When the result of the index test is used in establishing the final diagnosis.	1(0)	0	1(0)
	Observer variation (variation)	The reproducibility of test results is one of the determinants of diagnostic accuracy of an index test. Because of variation in observers, a test may not consistently yield the same result when repeated. In two or more observations of the same entity, intra-observer variability arises when the same person gets different results, and inter-observer variability, when two or more people disagree.	11(7)	0	3(1)
	Instrument variation (variation)	The reproducibility of test results is one of the determinants of diagnostic accuracy of an index test. Because of variation in laboratory procedures a test may not consistently yield the same result when repeated.	1(0)	0	1(0)
Analysis:	Handling of uninterpretable test results (bias)	A diagnostic test can produce an uninterpretable result with varying frequency depending on the test. These problems are often not reported in test efficacy studies with the uninterpretable results simply removed from the analysis. This may lead to the biased assessment of the test characteristics.	0	0	3(2)
	Withdrawals	If drop-outs from the study are not random they may lead to biased estimates of test performance.	1(0)	0	0
	Arbitrary choice of threshold value	The selection of the threshold value for the index test that maximises the sensitivity and specificity of the test may lead to overoptimistic measures of test performance. The performance of this cut-off in an independent set of patients may not be the same as in the original study.	1(0)	2(0)	2(0)
	Sample size	Small studies may produce less accurate estimates of test performance than larger studies.	1(0)	0	1(0)

Table 5.2: Summary of individual study results

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
Population								
Aldberg(2004)(21)	R	Variable	25 studies that used overall accuracy as summary measure	Disease Prevalence	When prevalence is low, overall accuracy more closely resembles specificity; when prevalence is high, overall accuracy more closely resembles sensitivity.	na	Na	Associated
Bachmann(2009)(22)	*M*	*Stress ECG*	*580 patients who underwent coronary angiography*	*Demographic Features*	*Proportion of patients with atypical symptoms*	na	na	*Associated*
Barber(2006)(23)	DA	Simple screening question for pelvic organ prolapse	120 women with high risk and 448 women at low risk	Disease Prevalence	High pre-test probability population	↑	↓	na
Biesheuvel(2008)(24)	*M*	*Tests for DVT*	*1295 consecutive patients with possible having deep vein thrombosis (DVT).*	*Distorted Selection of participants*	*Estimates from nested CC versus estimates from total cohort*	*na*	*na*	*None*
Boyer(2009)(25)	R	Diagnostic tests for carpel tunnel syndrome (CTS).	23 studies	Distorted Selection of participants	Use of case-control design (present in 14/23 studies)	↑	↑	↑
Burch(2006)(26)	R	Faecal occult blood tests (FOBT) in the detection of neoplasms	33 primary studies	Distorted Selection of participants	Case-control vs. cohort study	↑	na	na
Clark(2004)(27)	R	Tests for predicting endometrial hyperplasia	27 studies	Distorted Selection of participants	At least one of the following: adequate recruitment, appropriate spectrum, or adequate blinding	na	na	↓
Curtin (1997)(28)	DA	Body mass index (BMI)	226 Caucasians	Demographic features	Increased weight, being female	↑	none	na
Detrano (1988)(29) Detrano (1988)(30)	R	Exercise thallium scintigraphy	56 primary studies	Demographic features	Sex Age, medication use	associated	none	na
				Distorted selection of participants	Avoidance of limited challenge group	none	none	
				Disease severity	Inclusion of patients with prior myocardial infarction	↑	none	na
Detrano (1989)(31)	R	Exercise electro-cardiography	60 primary studies	Demographic features	Various patient related characteristics: not all associated	associated	associated	na
DiMatteo(2001)(32)	DA	Rapid antigen test	498 consecutive adults	Disease Severity	Increasing Centor criteria	↑	na	na
Egglin (1996)(33)	E	Pulmonary arteriography	24 arteriograms	Disease prevalence	Context of interpretation: effect of increased disease prevalence	↑	none	na
Elie(2008)(34)	DA	Papanicolaou smear test	1781 Women	Demographic Features	Age >35 years	None	↓	na

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
					Menopausal status, type of contraception, European origin, educational level, smoking	None	None	na
				Prior testing	Positive test for HPV	↑	↓	na
				Disease Prevalence	Referral setting vs. screening	↑	↓	na
Gaffkin(2010)(35)	DA	Visual inspection with acetic acid (VIA)	2182 women	Demographic Features	History of sexually transmitted diseases	None	None	na
				Prior Testing	Pap test status	None	None	na
Geleijnse(2009)(36)	R	Dobutamine stress echocardiography	62 studies	Demographic Features	History of MI	↑	None	na
					Medication use, age, gender	None	None	na
				Disease Severity	Extent of CAD (multivessel vs. single vessel involvement)	↑	None	na
				Distorted Selection of participants	Pre-test CAD probability	↑	↓	na
					Inclusion of patients with rest wall motion abnormalities	No effect	No effect	na
Gilbert(2002)(37)	R	EEG	25 studies	Demographic Features	Proportion of remote symptomatic patients, proportion of treated patients,	na	na	None
				Disease Prevalence	Sample probability of seizure recurrence	na	na	None
Haines(2007)(38)	R	Hospital fall risk screening tool	35 studies reporting 51 evaluations	Distorted Selection of participants	Retrospective vs. Prospective. Non-standard definition of prospective: In addition to the typical definition, an a priori defined cut-off was required to be classified as prospective.	na	na	↑
Hall(2004)(39)	DA	Rapid antigen detection test	561 children evaluated for pharyngitis.	Disease Severity	Increasing Centor criteria	↑	None	na
Hlatky (1984)(40)	DA	Exercise electro-cardiography	2269 patients	Demographic features	Exercise heart rate, number of disease arteries, type of angina, age and sex	associated	associated	na
Kittler(2002)(41)	R	Melanoma diagnosis with and without dermoscopy	27 studies	Disease Prevalence	Increased prevalence	na	na	↓
Lachs (1992)(42)	DA	Dipsticks	366 consecutive patients	Disease prevalence	High pre-test probability of disease	↑	↓	na
Leeflang(2009)(43)	M	Theoretical discussion illustrated with examples		Disease Prevalence	Direction and magnitude of effect varied across studies	Associated	Associated	na
Levy (1990)(44)	DA	Electro-cardiography	4684 patients with suspected left ventricular hypotrophy.	Demographic features	Sex (male), increased age, decreased BMI, not smoking	↑	none	na
				Disease severity	Increased severity of left ventricular hypertrophy	↑	none	na
Lijmer (1999)(18)	MR	Various different tests	184 primary studies of 218 tests	Distorted selection of participants	Diagnostic case-control studies	na	na	↑
					Non-consecutive patient enrolment	na	na	none
					Retrospective study design	na	na	none
					Failure to describe patient spectrum	na	na	↑
Mastandrea(20	R	BNP	67 studies (98 samples)	Demographic Features	Age, sex, BMI	na	na	None

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
08)(45)				Disease Severity	Disease severity	na	na	Associated
				Disease Prevalence	Disease prevalence	na	na	Associated
Medeiros(2007)(46)	DA	Confocal scanning laser opthalmoscopy (CSLO) in glaucoma.	Analysis 1: 67 eyes with visual field loss and 56 eyes of normal volunteers. Analysis 2: 83 suspected glaucoma	Distorted Selection of participants	Case-Control versus retrospective cohort - effect on AUC	na	na	↑
Melbye (1993)(47)	DA	Clinical cues	581 patients with suspected pneumonia	Disease prevalence	Increased prevalence	↑	↓	na
Michaud(2002)(48)	R	Various diagnostic tests for ventilator-associated pneumonia.	26 studies	Demographic Features	Prior treatment with antibiotics	Associated	Associated	na
				Distorted Selection of participants	Appropriate patient selection	↑	↑	na
Miller(2002)(49)	DA	SPECT	14 273 patients without known coronary artery disease	Demographic Features	Gender	None	None	na
Moons (1997)(50)	DA	Exercise test	295 consecutive patients with heart pain.	Demographic features Disease severity	Sex, workload, diabetes, smoking, cholesterol level (not all associated) Number of diseased vessels	↑ ↑	↓ none	na na
Morise (1994) (1995) (51;52)	DA	Exercise electro-cardiography	4467 patients with suspected coronary disease	Demographic factors	Men	↑	↑	na
O'Connor (1996)(53)	DA	Magnetic resonance imaging and evoked potentials	303 patients with suspected multiple sclerosis	Disease prevalence	Increased prevalence	↑	none	na
Philbrick (1982)(54)	DA	Graded exercise test	208 consecutive patients evaluated for coronary arterial disease	Distorted selection of participants	Exclusion of patients with other clinical conditions	na	na	↑
Pretorius(2007)(55)	DA	Acetic acid-aided visual inspection (VIA)	375 women with high-risk HPV or abnormal cervical cytology	Disease Severity	More severe disease	↑	na	na
Punglia(2003)(56)	M	PSA	6691 men	Demographic Features	Age (> vs. <60 years). Previous test results (abnormal DRE examination) showed no effect on accuracy after correcting for verification bias.	na	na	↓
Ransohoff (1978)(57)	R	Carcinoembryonic antigen (CEA) and nitro-blue tetrazolim (NBT) tests	17 studies of CEA and 16 of NBT	Disease severity	Extensive disease	↑	none	na
Roger (1997)(58)	DA	Exercise echocardiography	3679 consecutive patients	Demographic features	Men	↑	none	na
Rozanski (1983)(59)	DA	Exercise radionuclide ventriculoraphy	77 angio-graphically normal patients	Disease prevalence	Increased prevalence	not reported	↓	na

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
Rutjes(2006)(60)	MR	Various topics	31 meta-analyses (487 primary studies)	Distorted Selection of participants	Case-control design; Use/avoidance of limited challenge group; random vs. consecutive sampling	na	na	None
					Retrospective data collection Increased accuracy	na	na	↑
					Selection based on referral for index test results decreased accuracy	na	na	↓
Rutjes(2003)(61)	MR	Variable	49 meta-analyses (705 primary studies)	Distorted Selection of participants	Case-control design	None	None	None
					Retrospective design	None	None	None
					Consecutive enrolment	None	None	None
Santana-Boado (1998)(62)	DA	SPECT	702 consecutive patients evaluated for coronary disease	Demographic features	Sex	none	none	na
Shoaibi(2009)(63)	DA	Cardiac troponin (I (cTnI) assay	924 patients with possible myocardial ischemia	Demographic Features	Gender	None	None	na
Sohler(2008)(64)	DA	Psychiatric hospital diagnosis	491 psychiatric patients assigned final diagnoses	Demographic Features	Estimates of accuracy in black vs. white patients	None	None	na
Stein (1993)(65)	DA	Ventilation/ perfusion scan	1050 patients	Disease severity	Prior pulmonary disease	↑	none	na
Steinbauer (1998)(66)	DA	Screening tests for alcohol abuse	1333 adult family practice patients	Demographic features	Race and sex	na	na	associated
Stengel(2005)(67)	R	Ultrasonography	62 studies	Demographic Features	General population vs. children	↑	↑	na
					Penetrating versus non penetrating injuries.	None	None	na
				Disease Severity	Mean injury severity score	None	None	na
				Distorted Selection of participants	Reporting of selection criteria; consecutive enrolment; prospective design	None	None	na
Syed(2008)(68)	DA	PET MPI	833 PET studies performed in 122 patients without known CAD	Demographic Features	Female	↓	↑	na
					Obese	↓	↓	na
Taube (1990)(69)	M & DA	Tests for epithelial ovarian cancer	168 ovarian carcinoma patients	Disease severity	Clearly malignancy cases	↑	not reported	na
Thompson(2006)(70)	DA	PSA	5112 men on placebo; 4579 men finasteride	Demographic Features	Accuracy in men taking finasteride compared to men taking placebo.	↑	na	↑
Tobin(2006)(71)	R	Frequency-to-tidal volume ratio (f/Vt) in predicting weaning success.	29 Studies	Disease prevalence	Increasing prevalence increases the positive predictive value and decreases the negative predictive value	na	na	Associated
van der Schouw (1995)(72)	DA	Ultrasound	483 consecutive patients, 372 included	Disease prevalence	Increased prevalence (inclusion criteria widened)	↑	↑	na

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
Van Rijkom (1995)(73)	R	Tests for approximal caries	39 sets of sensitivity and specificity data	Distorted selection of participants	In vivo studies compared to in vitro studies	na	na	↑
Yoon(2009)(74)	DA	Myocardial perfusion imaging (MPI)	555 patients	Demographic Features	Beta-blocker therapy versus no beta-blocker therapy	None	None	na
Zhang(2002)(75)	DA	Routine ultrasound	Screening pregnant women; 3633 malformed foetuses	Disease Severity	Increased severity	↑	na	na
				Disease Prevalence	Increased prevalence of CHD or VSD	↓	na	na
Test protocol: materials and methods of testing								
Clark(2004)(27)	R	Tests for predicting endometrial hyperplasia	27 studies	Disease Progression	Delayed verification	na	na	↓
Davey(2006)(76)	R	Liquid-based cytology	56 studies	Test Technology	Liquid based cytology compared to conventional cytology	None	None	None
Detrano (1988)(29)	R	Exercise electro-cardiography	60 primary studies	Test execution Test technology Disease progression bias	Exercise protocol Automation of test Maximum interval between scintigraphy and angiography	none ↑ none	none ↓ none	na na na
Froelicher (1998)(77)	DA	Electrocardriography and angiographic callipers	814 consecutive patients with angina pectoris	Test technology	Computerised readings	none	none	na
Geleijnse(2009)(36)	R	Dobutamine stress echocardiography	62 studies	Test execution	Quantitative scoring of CAG	None	None	na
				Test Technology	Older vs. newer technology	None	None	na
Lijmer (1999)(18)	MR	Various different tests	184 primary studies of 218 tests	Test execution	Failure to describe index test execution Failure to describe reference standard execution	na na	na na	↑ ↓
Michaud(2002)(48)	R	Various diagnostic tests for ventilator-associated pneumonia.	26 studies	Test Technology	Higher BAL volume	↑	↑	na
Miller(2002)(49)	DA	SPECT	14 273 patients without known coronary artery disease	Test Technology	Type of radio-isotope technique	na	None	na
Rutjes(2006)(60)	MR	Various topics	31 meta-analyses (487 primary studies)	Disease Progression	Effect of time interval	na	na	None
				Treatment Paradox	Effect of treatment	na	na	None
Sonad(2001)(78)	R	MRI	27 studies	Test Technology	Fast SE imaging, <1.5T, non-endorectal coil	na	na	↑
Stengel(2005)(6 7)	R	Ultrasonography	62 studies	Test execution	Reporting of methods of test execution (no effect on sens), fast vs. fast+ US (no effect for sens or spec)	None	None	na
				Test Technology	Higher transducer frequency	↑	na	na
				Disease Progression	Reporting of time interval was associated with sensitivity; use of sufficiently short time interval showed no association	None	na	
Reference standard and verification procedure								

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
Arana (1990)(79)	R	Thyrotropin releasing hormone stimulation	10 studies	Inappropriate reference standard	Use of the DSM-III as opposed to the RDC as the reference standard	↓	not reported	na
Bowler (1998)(80)	DA	Necropsy	307 patients	Differential and partial verification bias	Necropsy to confirm the clinical diagnosis	na	na	"Scope for bias"
Boyer(2009)(81)	R	Diagnostic tests for carpel tunnel syndrome (CTS).	23 studies	Differential verification	Differential verification bias (present 4/23 studies)	None	None	None
Boyko (1988)(82)	M	Na	Formulas used to model theoretical effects	Inappropriate reference standard	Effects of reference standard errors	na	na	associated
Brealey(2007)(83)	R	Plain radiograph reading methods with radiography as reference standard	10 studies	Inappropriate reference standard	Use of less valid reference standard:	na	na	None
				Partial verification	Application of reference standard depending on observer's opinion	na	na	None
				Differential verification	Use of different reference standards in same study	na	na	None
Cagle(2009)(84)	DA	Colposcopy and visual inspection with acetic acid (VIA).	1839 women who attended screening	Inappropriate reference standard	Use of expanded vs. standard colposcopy. No effects were seen on sens or spec in the valuation of LBC or hc2 with either the expanded or standard reference standard.	↓	None	na
Cecil (1996)(85)	DA	Stress SPECT thallium testing	4354 records selected from computerised database	Partial verification bias	Effect of partial verification bias (Begg's method(33))	↑	↓	na
De Neef (1987)(86)	M	New rapid antigen detection tests	Models used to vary reference standard accuracy	Inappropriate reference standard	Increased sensitivity of the reference standard	↑	large errors	na
Detrano (1988)(29;30)	R	Exercise thallium scintigraphy	56 primary studies	Inappropriate reference standard Partial verification bias	Tomographic imaging instead of angiography as reference test Presence of partial verification bias	↑ none	↑ ↑	na na
Detrano (1989)(31)	R	Exercise electrocardiography	60 primary studies	Inappropriate reference standard Partial verification bias	Exercise test thought to be superior in accuracy as reference standard Presence of partial verification bias	associated na	not reported na	na none
Diamond (1991)(87)	M	na	Series of computer simulations using Begg-Greenes method(33)	Partial verification bias	Presence of partial verification bias	↑	↓	na
Diamond (1992)(88)	M	na	Series of computer simulations using Bayes' theorem	Partial verification bias	Presence of partial verification bias	↑	↓	na
Froelicher (1998)(77)	DA	Electrocardiography and angiographic callipers	814 consecutive patients with angina	Partial verification bias	Presence of partial verification bias	↑	↓	na
Gaffkin(2010)(35)	DA	Visual inspection with acetic acid (VIA)	2182 women	Partial verification	Presence of verification bias	↓	↓	na
Geleijnse(2009)(36)	R	Dobutamine stress echocardiography	62 studies	Partial verification	Presence of referral (partial verification) bias	None	↓	na

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
Gilbert(2002)(37)	R	EEG	25 studies	Inappropriate reference standard	Years followed (reference standard consisted of clinical follow-up)	na	na	None
Gupta(2003)(89)	R	PSA	3 studies	Partial verification	Partial verification bias	↑	↓	na
				Differential verification	Effect of differential verification where unverified test negative results were included in 2x2 table as true negative results	↑	↑	na
Lauer(2007)(90)	M	PET	534 consecutive patients with suspected lung cancer	Partial verification	Impact of verification bias for cancer of any site; Impact of verification bias on PET for detection of mediastinal cancer: no association	↑	↓	na
Lijmer (1999)(18)	MR	Various different tests	184 primary studies of 218 tests	Differential verification bias	Studies that used different reference standard	na	na	↑
				Partial verification bias	Presence of partial verification bias	na	na	none
Lijmer (1996)(91)	DA	Non-invasive tests	464 consecutive patients with suspected disease	Partial verification bias	Presence of partial verification bias	na	na	↑
Mastandrea(2008)(45)	R	BNP	67 studies (98 samples)	Reference standard	Reference Method	na	na	Associated
				Instrument Variation				
Michaud(2002)(48)	R	Various diagnostic tests for ventilator-associated pneumonia.	26 studies	Inappropriate reference standard	Use of diagnostic consensus criteria as reference standard	None	None	na
Miller(2002)(49)	DA	SPECT	14 273 patients without known coronary artery disease	Partial verification	Impact of adjusting for verification bias using either method (results similar for both methods)	↓	↑	na
Miller (1998)(92)	DA	Stress imaging	15945 low risk patients	Partial verification bias	Presence of partial verification bias	↑	↓	na
Mol (1999)(93)	R	Nuchal translucency measurement	25 studies	Partial verification bias	Presence of partial verification bias	↑	↑	na
Morise (1994) (1995)(51;52)	DA	Exercise electro-cardiography	4467 patients with suspected coronary disease	Partial verification bias	Presence of partial verification bias	↑	↓	na
Panzer (1987)(94)	DA	Clinical findings	374 patients with stroke and focal deficits	Partial verification bias	Presence of partial verification bias	↑	↓	na
Phelps (1995)(95)	M	na	Monte Carlo studies	Inappropriate reference standard	Use of inaccurate “fuzzy” reference standard	na	na	associated
Philbrick (1982)(54)	DA	Graded exercise test	208 consecutive patients	Partial verification bias	Presence of partial verification bias	↑	↓	na
Philbrick(2003)(96)	R	d-dimer test.	6 studies	Inappropriate reference standard	Estimates based on thigh imaging alone (optimal reference standard) compared to combined imaging of thigh and calf (imperfect reference standard)	↑	↓	na

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
Pretorius(2007)(55)	DA	Acetic acid-aided visual inspection (VIA)	375 women with high-risk HPV or abnormal cervical cytology	Inappropriate reference standard	Use of suboptimum reference standard	↑	na	na
Punglia(2003)(56)	M	PSA	6691 men	Partial verification	Impact of adjusting for verification bias	↓	↑	↑
Ransohoff (1982)(97)	R	Serum ferritin	Two studies	Partial verification bias	Presence of partial verification bias	↑	not reported	na
Ransohoff (1978)(57)	R	Carcinoembryonic antigen (CEA) and nitro-blue tetrazolim (NBT) tests	17 studies of CEA and 16 of NBT	Partial verification bias	Presence of partial verification bias	↑	not reported	na
Roger (1997)(58)	DA	Exercise echocardiography	3679 consecutive patients	Partial verification bias	Presence of partial verification bias	↑	↓	na
Rozanski (1983)(59)	DA	Exercise ventriculoraphy	77 angio-graphically normal patients	Partial verification bias	Presence of partial verification bias	not reported	↓	na
Rutjes(2006)(60)	MR	Various topics	31 meta-analyses (487 primary studies)	Inappropriate reference standard	Single vs. composite reference standard.	na	na	None
				Partial verification	Partial verification bias	na	na	None
				Differential verification	Differential verification bias	na	na	None
Rutjes(2003)(61)	MR	Variable	49 meta-analyses (705 primary studies)	Partial verification	Partial verification	None	None	None
				Differential verification	Differential verification	None	↑	↑
Santana-Boado (1998)(62)	DA	SPECT	702 consecutive low risk patients	Partial verification bias	Presence of partial verification bias	none	none	na
Stengel(2005)(67)	R	Ultrasonography	62 studies	Inappropriate reference standard	Use of single reference standard and reporting reference standard execution	↓	na	na
				Partial verification	Independent verification	↓	na	na
				Differential verification	Proportion of CT scans; proportion of laparotomies and proportion of diagnostic peritoneal lavage procedures (no effect)	Associated	na	na
Syed(2008)(68)	M	PET MPI	833 PET studies performed in 122 patients without known CAD	*Partial verification*	*Uncorrected (presence of partial verification)*	↑	↓	na
Thibodeau (1981)(98)	M	na	Various statistical models	Inappropriate reference standard	Use of inaccurate reference standard	na	na	associated
Van Rijkom (1995)(73)	R	Tests for approximal caries	39 sets of sensitivity and specificity data	Inappropriate reference standard	Use of weak validation methods	na	na	↑
Zhou (1994)(99)	M & DA	na	429 patients	Partial verification bias	Presence of partial verification bias	na	na	associated
Interpretation (reading process)								

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
Bachmann(2009)(22)	*M*	*Stress ECG*	*580 patients who underwent coronary angiography*	*Clinical Review Bias*	*ECG performance after formal incorporation of age, sex, and symptomatologyy*	na	na	↑
Berbaum (1988)(100)	E	Radiography	40 radiographs examined with and without clinical info.	Clinical review bias	Availability of clinical information	↑	none	↑
Berbaum (1989)(101)	E	Radiography	40 radiographs examined by a group of radiologist and a group of orthopaedic surgeons	Observer variation	Difference between radiologists and orthopaedic surgeons	na	na	associated
Boyer(2009)(81)	R	Diagnostic tests for carpel tunnel syndrome (CTS).	23 studies	Review Bias	Test review bias (present 8/23 studies). Diagnostic review bias (presented 2/23 studies) - no effect.	↑	None	↑
Brealey(2007)(83)	R	Plain radiograph reading methods with radiography as reference standard	10 studies	Review Bias	Reference standard review bias; no effect for test review bias: none	na	na	↑
Ciccone (1992)(102)	E	Mammography	45 mammograms, 7 radiologists	Observer variation	Inter- and intra-observer variation	na	na	associated
Cohen (1987)(103)	E	Fine-needle aspiration biopsy	50 specimens examined by 5 observers	Observer variation	Effect of training and experience	↑	↑	na
Corley (1997)(104)	E	Histologic diagnosis of pneumonia	39 lung biopsy samples, 4 pathologists	Observer variation	Inter- and intra-observer variation	na	na	none
Cuaron (1980)(105)	E	Tc-99m-phosphate myocardial imaging	250 myocardial slides evaluated by 6 observers	Observer variation	Inter-observer variation	na	na	associated
Detrano (1988)(29;30)	R	Exercise thallium scintigraphy	56 primary studies	Review bias	Lack of blinding i.e. presence of review bias	↑	not reported	↑
Detrano (1989)(31)	R	Exercise electrocardiography	60 primary studies	Review bias	Lack of blinding i.e. presence of review bias	na	na	none
Doubilet (1981)(106)	E	Radiographs	8 test films 4 with suggestive, 4 non-suggestive history	Clinical review bias	Suggestive clinical history	↑	↓	na
Eldevick (1982)(107)	E	Myelography and computed tomography	107 patients, assessed with and without clinical history	Clinical review bias	Availability of clinical information	↑	↓	na
Elie(2008)(34)	DA	Papanicolaou smear test	1781 Women	Clinical Review Bias	Clinical reading vs. optimised interpretation (blinded to clinical info and context)	None	None	na
Elmore (1994)(108)	E	Mammography	150 mammograms , 10 radiologists	Observer variation	Inter-observer variation	na	na	associated

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
Elmore (1997)(109)	E	Mammography	100 radiographs, assessed with and without clinical history	Clinical review bias	Availability of clinical information	na	na	↑
Erly(2003)(110)	DA	Emergency CT scans	716 consecutive CT scans	Observer Variation	Radiologist vs. neuroradiologist	↓	None	na
Froelicher (1998)(77)	DA	electrocardriography and angiographic callipers	814 consecutive patients with angina	Clinical review bias	Availability of clinical information	↑	not reported	na
Geleijnse(2009)(36)	R	Dobutamine stress echocardiography	62 studies	Review Bias	Blind reading of reference standard or index test (was blinded in all but 5 studies)	None	None	na
Gilbert(2002)(37)	R	EEG	25 studies	Observer Variation	Threshold for interpreting a positive EEG	na	na	Associated
Good (1990)(111)	E	Chest radiography	247 radiographs assessed with and without clinical history	Clinical review bias	Availability of clinical information	na	na	none
Gupta(2003)(89)	R	PSA	3 studies	Incorporation	Effect of incorporation bias	↑	↓	na
Haines(2007)(38)	R	Hospital fall risk screening tool	35 studies reporting 51 evaluations	Review Bias	Staff blinding	na	na	None
Irwig(2006)(112)	E	Ultrasound	Women with breast symptoms	Clinical Review Bias	Interpretation of ultrasound with mammography on view	na	na	None
Kittler(2002)(41)	R	Melanoma diagnosis with and without dermoscopy	27 studies	Review Bias	Test review bias	na	na	None
				Observer Variation	Dermoscopy interpreted by expert vs. non-expert examiners; dermoscopy interpreted by group of 2 or more experts vs. single interpretation	na	na	↑
				Instrument Variation	Accuracy of dermoscopy for experimental studies that used presentation of slides, colour prints, or digital images than for clinical studies in which diagnosis was made face to face	na	na	↓
Lijmer (1999)(18)	MR	Various different tests	184 primary studies of 218 tests	Review bias	Lack of blinding i.e. presence of review bias	na	na	↑
Mastandrea(2008)(45)	R	BNP	67 studies (98 samples)	Instrument Variation	Laboratory method	na	na	None
Moore(2005)(113)	DA	MRI	560 patients	Observer variation	Physical therapists and orthopaedic surgeons compared to non-orthopaedic providers	na	na	↑
Potchen (1979)(114)	E	Chest radiography	3 groups of radiologists: different combinations of data	Clinical review bias	Availability of clinical information	↑	not reported	na

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
Raab (1995)(115)	E	Bronchial brush specimens	100 bronchial brush specimens examined by different observers	Observer variation	Inter-observer variation	na	na	associated
Raab (2000)(116)	E	Bronchial brush specimens	97 specimens, assessed with and without clinical information	Clinical review bias	Availability of clinical information	na	na	↑
Ransohoff (1978)(57)	R	Carcinoembryonic antigen (CEA) and nitro-blue tetrazolim (NBT) tests	17 studies of CEA and 16 of NBT	Review bias	Lack of blinding i.e. presence of review bias	↑	↑	na
Ronco (1996)(117)	E	Colpohostological and cytolgic screening	61 samples examined by cytologists and experts	Observer variation	Effect of training and experience (being an "expert")	↑	not reported	na
Rutjes(2006)(60)	MR	Various topics	31 meta-analyses (487 primary studies)	Review Bias	Double blinded	na	na	None
				Incorporation	Incorporation bias	na	na	na
Rutjes(2003)(61)	MR	Variable	49 meta-analyses (705 primary studies)	Review Bias	Blinding	None	None	None
Schreiber (1963)(118)	E	Chest radiography	100 chest films, assessed with and without clinical information	Clinical review bias	Availability of clinical information	↑	none	Na
Stengel(2005)(67)	R	Ultrasonography	62 studies	Review Bias	Blinding against US results. Blinding against reference standard did not influence results.	↓	na	Na
				Observer Variation	Specification of sonography expertise and type of operatory (radiologist vs. surgeon)	None	na	Na
van der Aa(2010)(119)	DA	Cystoscopy	448 patients	Review Bias	Diagnostic review bias	↑	na	na
Wardlaw(2005)(120)	R	CT signs	15 studies	Clinical Review Bias	Knowledge of symptoms vs. no knowledge	None	None	na
				Observer Variation	Experienced observers	None	None	na
Analysis								
Detrano (1989)(31)	R	Exercise electro-cardiography	60 primary studies	Handling of indeterminate results	Treatment of equivocal or non-diagnostic tests	na	na	none
Ewald(2006)(121)	*M*	*Simulated data sets*	*Simulated data sets*	*Threshold selection*	*Data -driven threshold compared to pre-specified threshold. Size of bias decreases with increasing sample size*	↑	↑	na
Haines(2007)(38)	R	Hospital fall risk screening tool	35 studies reporting 51 evaluations	Sample size	Sample size	na	na	None
Leeflang(2008)(122)	*M*	*Theoretical examples*	*Various examples*	*Threshold selection*	*Data driven optimisation of threshold overestimates accuracy. Magnitude of bias greater with smaller sample sizes.*	↑	↑	na
Mastandrea(2008)(45)	R	BNP	67 studies (98 samples)	Threshold selection	Threshold selected to maximise accuracy vs. other method of threshold selection	na	na	None

Study details	Design*	Index test	Study population	Source of bias/variation	Factors investigated	Effect on sensitivity	Effect on specificity	Effect on overall accuracy
Philbrick (1982)(54)	DA	Graded exercise test	208 consecutive patients	Handling of indeterminate results	Exclusion of unsatisfactory exercise test results	na	na	unclear
Rutjes(2006)(60)	MR	Various topics	31 meta-analyses (487 primary studies)	Threshold selection	Post hoc definition of threshold	na	na	None
Sonad(2001)(78)	R	MRI	27 studies	Distorted Selection of participants	Sample size <30	na	na	↑
Stengel(2005)(67)	R	Ultrasonography	62 studies	Indeterminate Results	Handling of indeterminate results	None	na	na
				Withdrawals	Reporting of number of excluded patients and reporting of number of drop-outs	↓	na	na
Thompson(2006)(70)	DA	PSA	4579 men on placebo; 5112 on finasteride	Threshold selection	Fixed specificity in finasteride versus placebo arm	↑	na	na

* DA = diagnostic accuracy; R = review; E = experimental; M = modelling. Shaded rows depict studies included in the original bias review.

5.4 Summary of results

We classified sources of bias and/or variation for which there were at least 10 studies providing empirical evidence of bias as “considerable evidence”. Sources of bias/and or variation supported by 5 to 10 studies providing empirical evidence of bias were classed as “adequate evidence” and those supported by at least one but less than 5 studies were classed as “some evidence”. There was considerable evidence for the effects of demographic features, distorted selection of participants, disease prevalence, disease severity, inappropriate reference standard, partial verification bias, and observer variation. There was adequate evidence for the effects of differential verification bias, review bias, and clinical review bias. There was some evidence for the effects of prior testing, test technology, test execution, disease progression bias, incorporation bias, instrument variation, withdrawals, arbitrary choice of threshold and sample size. There was no evidence to support the effects of inappropriate handling of uninterpretable test results or treatment paradox on estimates of test performance.

5.5 Implications for QUADAS-2

Table 5.3 provides a summary of the evidence for each QUADAS item using the evidence rating outlines above.

Table 5.3 Summary of evidence of bias and/or variation by QUADAS item

Item		Source of bias and/or variation	Strength of evidence
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	Demographic features (variation) Distorted selection of participants (variation) Disease prevalence (variation) Disease severity (variation) Prior testing (variation)	Considerable
2.	Were selection criteria clearly described?	NA	None
3.	Is the reference standard likely to correctly classify the target condition?	Inappropriate reference standard	Considerable
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 (bias)	Some
5.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	Partial verification bias (bias)	Considerable
6.	Did patients receive the same reference standard regardless of the index test result?	Differential verification bias (bias)	Adequate
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 (bias)	Some
8.	Was the execution of the index test described in sufficient detail to permit replication of the test?	Test execution (variation)	Some
9.	Was the execution of the reference standard described in sufficient detail to permit its replication?	NA	None
10.	Were the index test results interpreted without knowledge of the results of the reference standard?	Review bias (bias)	Adequate
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?	Clinical review bias (bias)	Adequate
13.	Were uninterpretable/ intermediate test results reported?	Handling of uninterpretable test results (bias)	No evidence
14.	Were withdrawals from the study explained?	Withdrawals	Some
No equivalent QUADAS item:		Test technology (variation)	Some
		Treatment paradox (bias)	No evidence
		Observer variation (variation)	Considerable
		Instrument variation (variation)	Some
		Arbitrary choice of threshold value	Some
		Sample size	Some

Chapter 6: Review of studies that have evaluated QUADAS

Key Points

Eight studies reported evaluations of QUADAS. Three aimed to assess inter-rater reliability (IRR) and an additional study also provided data on IRR. This was a larger study conducted by us to evaluate QUADAS shortly after it was originally published. Three studies reported adaptations of QUADAS to particular situations. The final study assessed the fundamental mechanisms underlying spectrum and test review bias and the implications for QUADAS.

Studies were generally positive about QUADAS. Overall agreement in rating QUADAS items was generally good, but there was variation within items and across studies. Items 13 and 14 (reporting of uninterpretable results and withdrawals) consistently showed the lowest levels of agreement. One study highlighted problems with the item relating to availability of clinical information (12). The studies either modified QUADAS to include additional items for specific situations or recommended items for possible future inclusion in QUADAS.

Specific recommendations included:

- Consider modifying patient spectrum (item 1) by adding sub-categories
- Possible items for inclusion: extent to which index test represents current technology, observer details, explanation of withdrawals, appropriate statistical methods.
- Possible items for omission or clarification: reporting of uninterpretable results and withdrawals (13 and 14) and availability of clinical information (12).
- Consider expanding QUADAS to cover comparative tests and topics in which the reference standard consists of follow-up
- Emphasise importance of developing review specific scoring guidance, including specific items to be assessed

This review aims to provide a summary on the published data reporting reviewers' experience of using QUADAS.

6.1 Objectives

To evaluate all studies that have reported an evaluation of QUADAS.

6.2 Methods

We searched the following databases from inception to July 2010 using the term "QUADAS": MEDLINE, EMBASE, AMED, PsychINFO and CAB Abstracts. We included any study with the objective of evaluating QUADAS. This included studies reporting on inter-rater reliability and reviewers' opinions and experience of using QUADAS. Inclusion was assessed by one reviewer and checked by a second. One reviewer extracted data on the study objective, methods, positive and negative results in relation to QUADAS and any recommendations relating to the use or future development of QUADAS. Extraction was checked by a second reviewer. The results are presented stratified according to each QUADAS item.

6.3 Results

The literature searches produced 230 hits of which 12 appeared potentially relevant and full text copies were obtained. Eight of these studies fulfilled inclusion criteria and were included in the review. Three of the studies aimed to assess inter-rater reliability (IRR) when assessing QUADAS for particular topic areas (imaging and psychometric instruments),(123) and an additional study also provided data on IRR. This was a larger study conducted by us to evaluate QUADAS shortly after it was originally published.(2) Three studies reported adaptations of QUADAS to particular situations. One reported an adaptation of QUADAS to produce a new tool named "QUADRANOMICS" to address the methodological challenges posed by new molecular diagnostic test.(124) One assessed whether QUADAS captured all relevant sources of bias when a review involved comparative tests and when the reference standard involved longitudinal follow-up.(125) The third described modifications made to QUADAS to enable the assessment of diagnostic before-after studies and to describe experience using QUADAS.(126) The final study aimed to study and formalise the fundamental mechanisms underlying spectrum and test review bias and to suggest amendments to STARD and QUADAS based on this.(22)

General findings

Studies were generally positive about QUADAS with comments stating that it was informative, easy to use, allowed consistent and transparent rating, and authors stated that they would used it again. Criticisms related to the poor quality of reporting of the primary study which hampered quality assessment, the need for 2 papers to get the full QUADAS guidelines(1;2) and one of the studies stated that QUADAS did not lead to that much greater insight into the relationship between potential threats to validity identified by the checklist and the direction of results of the studies.

Item specific findings

Partial verification (Item 7): one study stated that they found this item difficult to score for case-control studies(127)

Availability of clinical information (Item 12): One study criticised the fact that QUADAS recommends recording contextual information when interpreting a test but does not stipulate how to use this information when assessing test performance. QUADAS recommends evaluating the index test using the same clinical data available when using the test in practice. This does not exclude the possibility of variation in index test performance when using different sets of clinical data as there could be different views on what clinical data should be used in test evaluation.

Inter-rater reliability

The overall agreement in rating QUADAS items was generally good, with the average agreement ranging from 69% to 90%, but there was greater variation within items and across studies. The only consistent finding across studies was that items 13 and 14 (reporting of uninterpretable results and withdrawals) showed the lowest levels of agreement with some of the studies reporting difficulties in applying the scoring guidelines to these items. Agreement was more variable across studies for other items. None of the items consistently showed good agreement across all studies. Agreement ranged from poor to moderate for items 2, 4 and 12 (description of selection criteria, time period, availability clinical data), from moderate to good for item 5, 6, 8 and 9 (partial and differential verification, description of index test and reference test) and from poor to high for items 1, 3, 7, 10 and 11 (spectrum, reference standard, incorporation, blinding index test and reference standard results).

Proposed additional items

The studies either included the following additional items or proposed that these should be included in QUADAS. Where items were recommended for particular topic areas this is indicated in brackets:

Spectrum composition:

- First indicate phase of study scale from 1 (healthy case-control study) to 4 (diagnostic cohort study).
- Additional details relating to patient spectrum:
    - duration of untreated disease
    - reason for referral of patients into the study
    - setting of the study
- Was the type of sample fully described?
- Were patients recruited consecutively?
- Was the study and/or collection of clinical variables conducted prospectively?

Test protocol: material and methods:

- Were the procedures and timing of biological sample collection with respect to clinical factors described with enough detail?
    - Clinical and physiological factors
    - Diagnostic and treatment procedures
- Were handling and pre-analytical procedures reported in sufficient detail and similar for the whole sample? If differences in procedure were reported was their effect on the results assessed?
- "Does the method used to perform the index test represent the current state of the art for that index test?". Similar wording for other items evaluating the clarity of reporting (items 2, 8, 9, 13 and 14) is suggested.
- Time between index test and reference standard: may not always be appropriate to have a short duration of follow-up

Interpretation:

- Items relating to mutual blinding of readers reviewing multiple tests (comparative tests).
- Who performed the clinical evaluation and image analysis? (imaging)

Analysis:

- Is it likely the presence of over fitting was avoided?
- What was the explanation for patients who did not receive CT or MRI? - sub question of item 14 (withdrawals)
- Do statistical method takes into account the lack of independence of results of index and comparator tests when derived from the same patients? (comparative tests)

Recommendations

Two studies mentioned the need for reviewers to provide clear guidance tailored to their review and to adhere to this guidance.(2;128) One of these, our earlier evaluation of QUADAS, recommended that reviewers should consider whether all QUADAS items are applicable to their review and whether additional quality items should be considered.(2) One study recommended that quality assessment be performed in duplicate.(128) One study suggested that future updates to QUADAS should consider additional criteria for situations in which a new index test is compared to a concurrent routine test and when the reference standard involves clinical follow-up.(125)

6.3 Summary

Studies were generally positive about QUADAS with authors stating that they would use QUADAS again. The overall agreement in rating QUADAS items was generally good, but there was variation within items and across studies. The only consistent finding across studies was that items 13 and 14 (reporting of uninterpretable results and withdrawals) showed the lowest levels of agreement. One study highlighted problems with the item relating to availability of clinical information (12). The studies either modified QUADAS to include additional items for specific situations or recommended items for possible future inclusion in QUADAS.

6.4 Implications for QUADAS-2

- Consider modifying patient spectrum (item 1) with the following possible sub-categories: study design, duration of untreated disease, reason for referral, setting, description of study sample, consecutive and/or prospective enrolment,
- Possible items for inclusion: extent to which index test represents current technology, observer details, explanation of withdrawals, appropriate statistical methods.

- Possible items for omission or clarification: reporting of uninterpretable results and withdrawals (13 and 14) and availability of clinical information (12).
- Consider expanding QUADAS to cover comparative tests and topics in which the reference standard consists of follow-up
- Emphasise importance of developing review specific scoring guidance, including specific items to be assessed

Chapter 7: Generating a list of items

7.1 Recommendations from the evidence base

The evidence provided by the reviews and survey undertaken suggests the following requirements for QUADAS in terms of general features and specific items for inclusion, modification, or exclusion.

General requirements for QUADAS-2 and accompanying guidance

- Emphasise importance of avoiding use of summary scores (Chapter 3, 4)
- Consider including explicit suggestions for overall rating of study quality and/or grouping studies based on quality (Chapter 3, 4)
- Emphasise importance of developing review specific scoring guidance (Chapter 4, 6)
- Consider including additional examples in the scoring guidance covering a broader variety of topics (Chapter 4)
- Consider providing an online learning resources that is continually updated based on reviewers' experience of using QUADAS (Chapter 4)
- Consider expanding QUADAS to cover the following situations:
    - Comparative tests (Chapter 4, 6)
    - Statistical correction for verification bias (Chapter 4)
    - Topics in which the reference standard consists of follow-up (Chapter 4, 6)
    - Remove items related to the quality of reporting (Chapter 4)

Content of QUADAS-2

Consider adding the following sub-categories relating to spectrum composition (Item 1):

- Study design (Chapter 3, 4, 6)
- Method of enrolment (Chapter 3, 6)
- Prospective/retrospective data collection (Chapter 4, 6)
- Unbiased patient selection (Chapter 4)
- Reporting patient selection (Chapter 6)
- Duration of untreated disease (Chapter 6)
- Reason for referral (Chapter 6)
- Setting (Chapter 6)

- Prior testing (Chapter 4)

Possible items for omission or clarification

- Incorporation bias (item 7) (Chapters 3, 4, 5)
- Availability of clinical review bias (item 12) (Chapters 3, 4, 6)
- Reporting of uninterpretable results (item 13) (Chapters 4, 5, 6)
- Reporting of withdrawals (item 14) (Chapter 4, 5, 6)

Possible items for inclusion:

- Test protocol
  - Treatment paradox (Chapter 3)
  - Test interpretation setting (Chapter 3)
  - Technological status of index test (Chapter 4, 6)
- Analysis
  - Inter-observer variability/experience (Chapter 3, 4, 5, 6)
  - Arbitrary choice of threshold value (Chapter 3)
  - Patient or segment unit of analysis (Chapter 3)
  - Reporting of methods of analysis (Chapter 3)
  - Appropriate methods of analysis (Chapter 4, 6)
  - Sample size (Chapter 4)
- Missing Data
  - Proportion of patients recruited enrolled (Chapter 3)
  - Explanation of withdrawals (Chapter 6)
- Other
  - Funding/Conflicts of interest (Chapters 3, 4)
  - Hypothesis (defined) (Chapter 4)

7.2 Conceptual decisions made by the steering group : factors that will affect the structure of QUADAS-2

Finalised conceptual decisions

- *Tool structure:* Restructure the tool to include two separate sections, one focusing on risk of bias and the other on applicability. Items relating to quality of reporting removed
- *Comparative tests:* Expand QUADAS-2 to cover this type of evaluation
- *Longitudinal follow-up:* QUADAS-2 will cover this type of evaluation
- *Prognostic/predictive tests:* QUADAS-2 will not cover predictive models
- *Topic specific items:* We will not broaden the scope of QUADAS to include topic-specific items either for test type (e.g. imaging, biochemistry), or clinical field.
- *Holistic nature of QUADAS-2:* We will aim to develop a set of independent criteria that work together, i.e. to ensure that there is no overlap between items.

Conceptual decisions open for discussion

- *Scoring:* Replace the scoring of "yes/no/unclear" with "high risk of bias" or "low risk of bias" following Cochrane structure. Separate the description of the basis for the scoring from the judgement of risk of bias. Consider how this can be adapted for the section of QUADAS-2 relating to applicability.
- *Sub items:* Add sub-items which will help to allow objective assessment of the key items.
- *Overall rating:* Consider including explicit suggestions for overall rating of study quality and/or grouping studies based on quality: can this be done, and if so how should we do this?

7.3 QUADAS-2

Table 7.1 summarise the evidence from each of the four phases of evidence gathering relating to each current QUADAS item, and suggested additional items, and highlights which items are proposed for inclusion, modification and exclusion based on these evaluations. We have suggested possible new items for inclusion if there was some evidence from Chapter 5 on their effects on accuracy measures and if at least one of the other Chapters proposed inclusion of this item. The table is colour coded so that original QUADAS items proposed for retention in QUADAS-2 are coloured green; items proposed for removal are coloured red; items suggested for removal but where discussion is required are coloured pink; items suggested for modification are coloured purple; and new items suggested for inclusion are coloured blue. Additional items suggested for inclusion in reviews assessing comparative tests are also listed.

Table 7.1 Summary of evidence from each of the four phases of evidence gathering for each current QUADAS item, and suggested additional items

a. Risk of Bias

Item		Source of bias and/or variation	Proposed QUADAS-2 Item	Recommended for inclusion/exclusion			Strength of evidence (Chapter 5)
				C3	C4	C6	
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	Distorted selection of participants	Were patients enrolled prospectively? Was a random or consecutive sample of patient enrolled? Did the study avoid using a case-control design?	✗ ✓ ✓	✓ ✗ ✓	✓ ✓ ✓	Considerable
2.	Were selection criteria clearly described?	NA		✗	✓	✗	None
3.	Is the reference standard likely to correctly classify the target condition?	Inappropriate reference standard	RETAIN ORIGINAL ITEM	na	na	na	Considerable
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	CONSIDER REMOVING AND HOW TO HANDLE FOR REFERENCE STANDARD THAT INCLUDES FOLLOW-UP	✗	✗	✗	Some
5.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	Partial verification bias	SHOULD WE INCLUDE SUB-QUESTIONS? HOW TO HANDLE STATISTICAL CORRECTION OF VERIFICATION BIAS	na	na	na	Considerable
6.	Did patients receive the same reference standard regardless of the index test result?	Differential verification bias		na	na	na	Adequate
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	SHOULD THIS BE RETAINED?	✓	✓	✗	Some
8.	Was the execution of the index test described in sufficient detail to permit replication of the test?	Test execution (variation)		✗	✓	✗	Some
9.	Was the execution of the reference standard described in sufficient detail to permit its replication?	NA		✗	✓	✗	None
10.	Were the index test results interpreted without knowledge of the results of the reference standard?	Review bias	RETAIN ORIGINAL ITEMS	na	na	na	Adequate
11.	Were the reference standard results interpreted without knowledge of the results of the index test?			na	na	na	
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	SHOULD THIS BE RETAINED?	✓	✓	✓	Adequate

13. Were uninterpretable/ intermediate test results reported?	Handling of uninterpretable test results	SHOULD THESE BE INCLUDED USING DIFFERENT PHRASING REGARDING WHETHER MISSING DATA AND/OR UNINTERPRETABLE TEST RESULTS WERE HANDLED ADEQUATELY?	✗	✓	✓	No evidence
14. Were withdrawals from the study explained?	Withdrawals		✗	✓	✓	Some
Possible additional items	Treatment Paradox	Was treatment started after the index test results were available prior to confirmation of the diagnosis with the reference standard?	✓	✗	✗	Some
	Arbitrary choice of threshold value	Was the threshold derived independently of the results of the study?	✓	✗	✗	Some
	Sample size	Did the study include an adequate sample size? IF WE WANT TO INCLUDE THIS, WHAT IS CONSIDERED ADEQUATE?	✗	✓	✗	Some

C3=Chapter 3; C4=Chapter 4; C6=Chapter 6; ✓=include; ✗=exclude

b. Applicability

Item	Source of variation	Proposed QUADAS Item	Recommended for inclusion			Strength of evidence
			C3	C4	C6	
1. Was the spectrum of patients representative of the patients who will receive the test in practice?	Demographic features Disease prevalence Disease severity Prior testing	Were the following consistent with the intended use of the index test? Reason for referral Setting Prior testing	✗ ✗ ✗	✗ ✗ ✓	✓ ✓ ✗	Considerable
Possible additional items	Test technology	Was the technology of the index test current?	✗	✓	✓	Adequate
Possible additional items	Test interpretation setting	Was the test interpreted in the same setting as it would be in practice?	✓	✗	✗	Adequate
Possible additional items	Observer variation	Was the test interpreted by someone with the same level of expertise who would interpret the test in practice?	✓	✓	✓	Considerable

C3=Chapter 3; C4=Chapter 4; C6=Chapter 6; ✓=include; ✗=exclude

Items relating to comparative tests

We suggest that for reviews assessing comparative test, the following additional items are added. These are all based on existing QUADAS items

- Did the whole sample or a random selection of the sample, undergo both the index test and the comparator?
- Is the time period between the index test and the comparator test short enough to be reasonably sure that the target condition did not change between the two tests?

- Were the results of both the index test and the comparator test verified with the same reference standard?
- Was the reference standard independent of the comparator test (i.e. the comparator tests did not form part of the reference standard)?
- Were the results of the index test interpreted without knowledge of the comparator test results?
- Did the same number of uninterpretable / intermediate test results occur for the index test as for the comparator test?
- Something about patients who do receive test A, get lost before they receive test B and/or (again) get lost before they receive the reference standard?

8. References

(1) Whiting P, Rutjes A, Reitsma J, Bossuyt P, 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) Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Medical Research Methodology 2006; 6:9.

(3) Reitsma JB, Rutjes AWS, WP, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration; 2009.

(4) Moher D, Schulz KF, Simera I, Altman DG. Providing Guidance to Developers of Health Reporting Guidelines. In press 2009.

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

(6) Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. J Clin Epidemiol 2005; 58(1):1-12.

(7) West S, King V, Carey T, Lohr K, McKoy N, Sutton S et al. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47; AHRQ Publication No. 02-E016. 2002. University of North Carolina: Agency for Healthcare Research and Quality.

(8) Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002; 2:9.

(9) Akcil M, Karaagaoglu E, Demirhan B. Diagnostic accuracy of fine-needle aspiration cytology of palpable breast masses:an SROC curve with fixed and random effects linear meta-regression models. Diagnostic Cytopathology 2008;303-310.

(10) Ngamruengphong S, Sharma VK, Das A. Diagnostic yield of methylene blue chromo endoscopyf or detecting specialized intestinal metaplasia and dysplasia in Barrett's esophagus : a meta-analysis. Gastrointestinal Endoscopy 2009;1021-1028.

(11) Tandon S, Shahab R, Benton JI, Ghosh SK, Sheard J, Jones TM. Fine-needle aspiration cytology in a regional head and neck cancer center : comparison with a systematic review and meta-analysis. Head and Neck 2008;1246-1252.

(12) Sackett DL, Staus SE, Richardson WS. Evidence based medicine: how to practice and teach EBM. 2nd ed. New York: Churchill Livingstone; 2000.

(13) NHS Public Health Resource Unit. 12 questions to help you make sense of a diagnostic test study. http://www phru nhs uk/Pages?PHD/resources htm [2008 [cited 2008 July 1];

(14) Banal F, Dougados M, Combescure C, Gossec L. Sensitivity and specificity of the American College of Rheumatology 1987 criteria for the diagnosis of rheumatoid arthritis according to disease duration : a systematic literature review and meta-analysis. Annals of the Rheumatic Diseases 2009;1184-1191.

(15) NHS Centre for Reviews and Dissemination. Undertaking systematic reviews of research on effectiveness. CRD Report No. 4. 2001. York, University of York.

(16) Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC Health Services Research 2002; 2(1):4.

(17) Kelly S, Berry E, Roderick P, Harris KM, Cullingworth J, Gathercole L et al. The identification of bias in studies of the diagnostic performance of imaging modalities. Br J Radiol 1997; 70(838):1028-1035.

(18) Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11):1061-1066.

(19) McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. JAMA 2000; 284(1):79-84.

(20) Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987; 6(4):411-423.

(21) Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. Journal of General Internal Medicine 2004; 19(5):460-465.

(22) Bachmann LM, ter RG, Weber WE, Kessels AG. Multivariable adjustments counteract spectrum and test review bias in accuracy studies. Journal of Clinical Epidemiology 2009; 62(4):357-361.

(23) Barber MD, Neubauer NL, Klein-Olarte V. Can we screen for pelvic organ prolapse without a physical examination in epidemiologic studies? American Journal of Obstetrics & Gynecology 2006; 195(4):942-948.

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

(25) Boyer K, Wies J, Turkelson CM. Effects of bias on the results of diagnostic studies of carpal tunnel syndrome. [Review] [41 refs]. Journal of Hand Surgery - American Volume 2009; 34(6):1006-1013.

(26) Burch J, Westwood M, Soares-Weiser K. Should data from case-controlled studies be included in systematic reviews alongside diagnostic cohort studies? [abstract]. XIV Cochrane Colloquium; 2006 October 23 26; Dublin, Ireland 2006;86.

(27) Clark TJ, ter RG, Coomarasamy A, Khan KS. Bias associated with delayed verification in test accuracy studies: accuracy of tests for endometrial hyperplasia may be much higher than we think!. [Review] [36 refs]. BMC Medicine 2004; 2:18.

(28) Curtin F, Morabia A, Pichard C, Slosman DO. Body mass index compared to dual-energy x-ray absorptiometry: evidence for a spectrum bias. J Clin Epidemiol 1997; 50(7):837-43.

(29) Detrano R, Janosi A, Lyons KP, Marcondes G, Abbassi N, Froelicher VF. Factors affecting sensitivity and specificity of a diagnostic test: the exercise thallium scintigram. Am J Med 1988; 84(4):699-710.

(30) Detrano R, Lyons KP, Marcondes G, Abbassi N, Froelicher VF, Janosi A. Methodologic problems in exercise testing research. Are we solving them? Arch Intern Med 1988; 148(6):1289-95.

(31) Detrano R, Gianrossi R, Mulvihill D, Lehmann K, Dubach P, et al. Exercise-induced ST segment depression in the diagnosis of multivessel coronary disease: a meta analysis. JACC 1989; 14(6):1501-8.

(32) Dimatteo LA, Lowenstein SR, Brimhall B, Reiquam W, Gonzales R. The relationship between the clinical features of pharyngitis and the sensitivity of a rapid antigen test: evidence of spectrum bias. Annals of Emergency Medicine 2001; 38(6):648-652.

(33) Egglin TKP, Feinstein AR. Context bias: a problem in diagnostic radiology. JAMA 1996; 276(21):1752-1755.

(34) Elie C, Coste J, French Society of Clinical Cytology Study Group. A methodological framework to distinguish spectrum effects from spectrum biases and to assess diagnostic and screening test accuracy for patient populations: application to the Papanicolaou cervical cancer smear test. BMC Medical Research Methodology 2008; 8:7.

(35) Gaffikin L, McGrath J, Arbyn M, Blumenthal PD. Avoiding verification bias in screening test evaluation in resource poor settings: a case study from Zimbabwe. Clinical Trials 2008; 5(5):496-503.

(36) Geleijnse ML, Krenning BJ, van Dalen BM, Nemes A, Soliman OI, Bosch JG et al. Factors affecting sensitivity and specificity of diagnostic testing: dobutamine stress echocardiography. [Review] [82 refs]. Journal of the American Society of Echocardiography 2009; 22(11):1199-1208.

(37) Gilbert DL, Sethuraman G, Kotagal U, Buncher CR. Meta-analysis of EEG test performance shows wide variation among studies (Provisional abstract). Neurology 2003; 60(4):564-570.

(38) Haines TP, Hill K, Walsh W, Osborne R. Design-related bias in hospital fall risk screening tool predictive accuracy evaluations: systematic review and meta-analysis. [Review] [70 refs]. Journals of Gerontology Series A-Biological Sciences & Medical Sciences 2007; 62(6):664-672.

(39) Hall MC, Kieke B, Gonzales R, Belongia EA. Spectrum bias of a rapid antigen detection test for group A beta-hemolytic streptococcal pharyngitis in a pediatric population. Pediatrics 2004; 114(1):182-186.

(40) Hlatky MA, Pryor DB, Harrell FE, Jr., Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984; 77(1):64-71.

(41) Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. Lancet Oncology 2002; 3(3):159-165.

(42) Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med 1992; 117:135-140.

(43) Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. [Review] [38 refs]. Journal of Clinical Epidemiology 2009; 62(1):5-12.

(44) Levy D, Labib S, Anderson K. Determinant of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. Circulation 1990; 81:815-820.

(45) Mastandrea P. Some heterogeneity factors affecting the B-type natriuretic peptides outcome: a meta-analysis. [Review] [77 refs]. Clinical Chemistry & Laboratory Medicine 2008; 46(12):1687-1695.

(46) Medeiros FA, Ng D, Zangwill LM, Sample PA, Bowd C, Weinreb RN. The effects of study design and spectrum bias on the evaluation of diagnostic accuracy of confocal scanning laser ophthalmoscopy in glaucoma. Investigative Ophthalmology & Visual Science 2007; 48(1):214-222.

(47) Melbye H, Straume B. The spectrum of patients strongly influences the usefulness of diagnostic tests for pneumonia. Scand J Prim Health Care 1993; 11(4):241-6.

(48) Michaud S, Suzuki S, Harbarth S. Effect of design-related bias in studies of diagnostic tests for ventilator-associated pneumonia. [Review] [50 refs]. American Journal of Respiratory & Critical Care Medicine 2002; 166(10):1320-1325.

(49) Miller TD, Hodge DO, Christian TF, Milavetz JJ, Bailey KR, Gibbons RJ. Effects of adjustment for referral bias on the sensitivity and specificity of single photon emission computed tomography for the diagnosis of coronary artery disease. American Journal of Medicine 2002; 112(4):290-297.

(50) Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997; 8(1):12-7.

(51) Morise AP, Diamond GA. Comparison of the sensitivity and specificity of exercise electrocardiography in biased and unbiased populations of men and women. Am Heart J 1995; 130(4):741-7.

(52) Morise AP, Diamond GA. Does sex discrimination explain the differences in test accuracy among men and women referred for exercise electrocardiography? 67th Scientific Sessions of the American Heart Association, Dallas, Texas, USA, November 1994; 90(4 PART 2).

(53) O'Connor PW, Tansay CM, Detsky AS, Mushlin AI, Kucharczyk W. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. Neurology 1996; 47(1):140-4.

(54) Philbrick JT, Horwitz RI, Feinstein AR, Langou RA, Chandler JP. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. JAMA 1982; 248(19):2467-70.

(55) Pretorius RG, Bao YP, Belinson JL, Burchette RJ, Smith JS, Qiao YL. Inappropriate gold standard bias in cervical cancer screening studies. International Journal of Cancer 2007; 121(10):2218-2224.

(56) Punglia RS, D'Amico AV, Catalona WJ, Roehl KA, Kuntz KM. Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen. New England Journal of Medicine 2003; 349(4):335-342.

(57) Ransohoff D, Feinstein A. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978; 299:926-929.

(58) Roger VL, Pellikka PA, Bell MR, Chow CW, Bailey KR, Seward JB. Sex and test verification bias. Impact on the diagnostic value of exercise echocardiography. Circulation 1997; 95(2):405-10.

(59) Rozanski A, Diamond GA, Berman D, Forrester JS, Morris D, Swan HJ. The declining specificity of exercise radionuclide ventriculography. N Engl J Med 1983; 309(9):518-22.

(60) Rutjes AW, Reitsma JB, Di NM, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ Canadian Medical Association Journal 2006; 174(4):469-476.

(61) Rutjes AWS, Smidt N, Di NM, Lijmer JG, Mol BWJ, van Rijn JC et al. Study design features affect estimates of sensitivity and specificity, but effects may vary. Sources of bias and variation in diagnostic accuracy studies, Thesis. University of Amsterdam, The Netherlands; 2005. 121-136.

(62) Santana-Boado C, Candell-Riera J, Castell-Conesa J, Aguade-Bruix S, Garcia-Burillo A, Canela T et al. Diagnostic accuracy of technetium-99m-MIBI myocardial SPECT in women and men. J Nucl Med 1998; 39(5):751-5.

(63) Shoaibi A, Tavris DR, McNulty S. Gender differences in correlates of troponin assay in diagnosis of myocardial infarction. Translational Research: The Journal Of Laboratory & Clinical Medicine 2009; 154(5):250-256.

(64) Sohler NL, Bromet EJ. Does racial bias influence psychiatric diagnoses assigned at first hospitalization? Social Psychiatry & Psychiatric Epidemiology 2003; 38(8):463-472.

(65) Stein PD, Gottschalk A, Henry JW, Shivkumar K. Stratification of patients according to prior cardiopulmonary disease and probability assessment based on the number of mismatched segmental equivalent perfusion defects. Approaches to strengthen the diagnostic value of ventilation/perfusion lung scans in acute pulmonary embolism. Chest 1993; 104(5):1461-7.

(66) Steinbauer JR, Cantor SB, Holzer CE, Volk RJ. Ethnic and sex bias in primary care screening tests for alcohol use disorders. Ann Intern Med 1998; 129(5):353-362.

(67) Stengel D, Bauwens K, Rademacher G, Mutze S, Ekkernkamp A. Association between compliance with methodological standards of diagnostic research and reported test accuracy: meta-analysis of focused assessment of US for trauma. Radiology 2005; 236(1):102-111.

(68) Syed IS, Miller T, Gibbons RJ, Askew JW, Hodge D, Chareonthaitawee P. Effect of Referral Bias on the Diagnostic Accuracy of N-13 Ammonia and Rubidium-82 Myocardial Perfusion Imaging with Positron Emission Tomography in the Detection of Coronary Artery Disease. Circulation 2008; 118(18, Suppl. 2).

(69) Taube A, Tholander B. Over- and underestimation of the sensitivity of a diagnostic malignancy test due to various selections of the study population. Acta Oncol 1990; 29:1-5.

(70) Thompson IM, Chi C, Ankerst DP, Goodman PJ, Tangen CM, Lippman SM et al. Effect of finasteride on the sensitivity of PSA for detecting prostate cancer. Journal of the National Cancer Institute 2006; 98(16):1128-1133.

(71) Tobin MJ, Jubran A. Variable performance of weaning-predictor tests: role of Bayes' theorem and spectrum and test-referral bias. [Review] [67 refs]. Intensive Care Medicine 2006; 32(12):2002-2012.

(72) van der Schouw YT, Van Dijk R, Verbeek AL. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. J Clin Epidemiol 1995; 48(3):417-22.

(73) van Rijkom HM, Verdonschot EH. Factors involved in validity measurements of diagnostic tests for approximal caries--a meta-analysis. Caries Research 1995; 29(5)):364-70.

(74) Yoon AJ, Melduni RM, Duncan SA, Ostfeld RJ, Travin MI. The effect of beta-blockers on the diagnostic accuracy of vasodilator pharmacologic SPECT myocardial perfusion imaging. Journal of Nuclear Cardiology 2009; 16(3):358-367.

(75) Zhang WH, Levi S, Alexander S, Viart P, Grandjean H, Eurofetus Study Group. Sensitivity of ultrasound screening for congenital anomalies in unselected pregnancies. Revue d Epidemiologie et de Sante Publique 2002; 50(6):571-580.

(76) Davey E, Barratt A, Irwig L, Chan SF, Macaskill P, Mannes P et al. Effect of study design and quality on unsatisfactory rates, cytology classifications, and accuracy in liquid-based versus conventional cervical cytology: a systematic review. Lancet 2006; 367(9505):122-132.

(77) Froelicher VF, Lehmann KG, Thomas R, Goldman S, Morrison D, Edson R et al. The electrocardiographic exercise test in a population with reduced workup bias: diagnostic performance, computerized interpretation, and multivariable prediction. Veterans Affairs Cooperative Study in Health Services :016 (QUEXTA) Study Group. Quantitative Exercise Testing and Angiography. Ann Intern Med 1998; 128(12 Pt 1):965-74.

(78) Sonnad SS, Langlotz CP, Schwartz JS. Accuracy of MR imaging for staging prostate cancer: a meta-analysis to examine the effect of technologic change. Acad Radiol 2001; 8(2):149-157.

(79) Arana GW, Zarzar MN, Baker E. The effect of diagnostic methodology on the sensitivity of the TRH stimulation test for depression: a literature review. Biological Psychiatry 1990; 28(8):733-7.

(80) Bowler JV, Munoz DG, Merskey H, Hachinski V. Fallacies in the pathological confirmation of the diagnosis of Alzheimer's disease. J Neurol Neurosurg Psychiatry 1998; 64(1):18-24.

(81) Boyer K, Wies J, Turkelson CM. Effects of bias on the results of diagnostic studies of carpal tunnel syndrome. [Review] [41 refs]. Journal of Hand Surgery - American Volume 2009; 34(6):1006-1013.

(82) Boyko EJ, Alderman BW, Baron AE. Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. J Gen Intern Med 1988; 3(5):476-81.

(83) Brealey SD, Scally AJ, Hahn S, Godfrey C. Evidence of reference standard related bias in studies of plain radiograph reading performance: a meta-regression. [Review] [48 refs]. British Journal of Radiology 2007; 80(954):406-413.

(84) Cagle A, Hu S, Sellors J, Bao Y, Lim J, Li S et al. Use of an expanded gold standard to estimate the accuracy of colposcopy and visual inspection with acetic acid. International Journal of Cancer 2010; 126(1):156-161.

(85) Cecil MP, Kosinski AS, Jones MT, Taylor A, Alazraki NP, Pettigrew RI et al. The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. J Clin Epidemiol 1996; 49(7):735-42.

(86) De Neef P. Evaluating rapid test for streptococcal pharyngitis: the apparent accuracy of a diagnostic test when there are errors in the standard of comparison. Med Decis Making 1987; 7:92-6.

(87) Diamond GA. Affirmative actions: can the discriminant accuracy of a test be determined in the face of selection bias? Med Decis Making 1991; 11(1):48-56.

(88) Diamond GA. Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem. Med Decis Making 1992; 12(1):22-31.

(89) Gupta A, Roehrborn CG. Verification and incorporation biases in studies assessing screening tests: prostate-specific antigen as an example. Urology 2004; 64(1):106-111.

(90) Lauer MS, Murthy SC, Blackstone EH, Okereke IC, Rice TW. [18F]Fluorodeoxyglucose uptake by positron emission tomography for diagnosis of suspected lung cancer: impact of verification bias. Archives of Internal Medicine 2007; 167(2):161-165.

(91) Lijmer JG, Hunink MG, van den Dungen JJ, Loonstra J, Smit AJ. ROC analysis of noninvasive tests for peripheral arterial disease. Ultrasound Med Biol 1996; 22(4):391-8.

(92) Miller TD, Hodge DO, Christian TF, Milavetz JJ, Bailey KR, Gibbons RJ. The impact of adjusting for post-test referral bias on apparent sensitivity and specificity of SPECT myocardial perfusion imaging in men and women. 47th Annual Scientific Session of the American College of Cardiology, Atlanta, Georgia, USA, March 1998; 31(2 SUPPL. A).

(93) Mol BW, Lijmer JG, van der Meulen J, Pajkrt E, Bilardo CM, Bossuyt PM. Effect of study design on the association between nuchal translucency measurement and Down syndrome. Obstet Gynecol 1999; 94(5 Pt 2):864-9.

(94) Panzer R, Suchman A, Griner P. Workup bias in prediction research. Med Decis Making 1987; 7:115-119.

(95) Phelps CE, Hutson A. Estimating diagnostic test accuracy using a "fuzzy gold standard". Med Decis Making 1995; 15(1):44-57.

(96) Philbrick JT, Heim S. The d-dimer test for deep venous thrombosis: gold standards and bias in negative predictive value. Clinical Chemistry 2003; 49(4):570-574.

(97) Ransohoff D, Muir W. Diagnostic work-up bias in the evalutaion of a test: serum ferritin and hereditary hemochromatosis. Med Decis Making 1982; 2:139-146.

(98) Thibodeau L. Evaluating diagnostic tests. Biometrics 1981;801-804.

(99) Zhou XH. Effect of verification bias on positive and negative predictive values. Stat Med 1994; 13(17):1737-45.

(100) Berbaum KS, el-Khoury GY, Franken EA, Jr., Kathol M, Montgomery WJ, Hesson W. Impact of clinical history on fracture detection with radiography. Radiology 1988; 168(2):507-11.

(101) Berbaum KS, Franken EA, Jr., el-Khoury GY. Impact of clinical history on radiographic detection of fractures: a comparison of radiologists and orthopedists. American Journal of Roentgenology 1989; 153(6):1221-4.

(102) Ciccone G, Vineis P, Frigerio A, Segnan N. Inter-observer and intra-observer varaibility of mammorgram interpretation: a field study. Eur J Cancer 1992; 28A:1054-1058.

(103) Cohen MB, Rodgers RPC, Hales MS, Gonzales JM, Ljung BME, Beckstead JH et al. Influence of training and experience in fine-needle aspiration biopsy of breast - receiver operating characteristics curve analysis. Arch Pathol Lab Med 1987; 111(6):518-520.

(104) Corley DE, Kirtland SH, Winterbauer RH, Hammar SP, Dail DH, Bauermeister DE et al. Reproducibility of the histologic diagnosis of pneumonia among a panel of four pathologists: analysis of a gold standard. Chest 1997; 112(2):458-65.

(105) Cuaron A, Acero AP, Cardenas M, et al. Interobserver variability in the interpretation of myocardial images with Tc-99m-labeled diphosphonate and pyrophosphate. J Nucl Med 1980; 21(1):1-9.

(106) Doubilet P, Herman P. Interpretation of radiographs: effect of clinical history. Am J Roentgenol 1981; 137:1055-1058.

(107) Eldevik O, Dugstad G, Orrison W, Haughton V. The effect of clinical bias on the interpretation of myelography and spinal computer tomography. Radiology 1982; 145:85-89.

(108) Elmore J, Wells C, Lee C. Variability in radiologists' interpretation of mammorgrams. New England Journal of Medicine 1994; 331:1493-1499.

(109) Elmore J, Wells C, Howard D. The impact of clinical history on mammographic interpretations. JAMA 1997; 277:49-52.

(110) Erly WK, Ashdown BC, Lucio RW, Carmody RF, Seeger JF, Alcala JN. Evaluation of emergency CT scans of the head: is there a community standard? AJR Am J Roentgenol 2003; 180(6):1727-1730.

(111) Good B, Cooperstein L, DeMarino G. Does knowledge of the clinical history affect the accuracy of chest radiograph interpretiation? Am J Roentgenol 1990; 154:709-712.

(112) Irwig L, Macaskill P, Walter SD, Houssami N. New methods give better estimates of changes in diagnostic accuracy when prior information is provided. Journal of Clinical Epidemiology 2006; 59(3):299-307.

(113) Moore JH, Goss DL, Baxter RE, DeBerardino TM, Mansfield LT, Fellows DW et al. Clinical diagnostic accuracy and magnetic resonance imaging of patients referred by physical therapists, orthopaedic surgeons, and nonorthopaedic providers. Journal of Orthopaedic & Sports Physical Therapy 2005; 35(2):67-71.

(114) Potchen E, Gard J, Lazar P, Lahaie P, Andary M. The effect of clinical history data on chest film interpretation: durection or istraction. Invest Radiol 1979; 14:404.

(115) Raab SS, Thomas PA, Lenel JC, Bottles K, Fitzsimmons KM, Zaleski MS et al. Pathology and probability: likelihood ratios and receiver operating characteristic curves in the interpretation of bronchial brush specimens. Am J Clin Pathol 1995; 103(5):588-593.

(116) Raab SS, Oweity T, Hughes JH, Salomao DR, Kelley CM, Flynn CM et al. Effect of clinical history on diagnostic accuracy in the cytologic interpretation of bronchial brush specimens. Am J Clin Pathol 2000; 114(1):78-83.

(117) Ronco G, Montanari G, Aimone V, Parisio F, Segnan N, Valle A et al. Estimating the sensitivity of cervical cytology: errors of interpretation and test limitations. Cytopathology 1996; 7(3):151-8.

(118) Schreiber M. The clinical history as a factor in roentgenogram interpretation. JAMA 1963; 185:137-139.

(119) van der Aa MN, Steyerberg EW, Bangma C, van Rhijn BW, Zwarthoff EC, van der Kwast TH. Cystoscopy revisited as the gold standard for detecting bladder cancer recurrence: diagnostic review bias in the randomized, prospective CEFUB trial. Journal of Urology 2010; 183(1):76-80.

(120) Wardlaw JM, Mielke O. Early signs of brain infarction at CT: observer reliability and outcome after thrombolytic treatment. Systematic review (Structured abstract). Radiology 2005; 235(2):444-453.

(121) Ewald B. Post hoc choice of cut points introduced bias to diagnostic research.[Erratum appears in J Clin Epidemiol. 2007 Jul;60(7):756]. Journal of Clinical Epidemiology 2006; 59(8):798-801.

(122) Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clinical Chemistry 2008; 54(4):729-737.

(123) Bauwens K, Ekkernkamp A, Stengel D. QUADAS: early experience with a new methodological scoring tool for diagnostic meta-analyses [abstract]. XIII Cochrane Colloquium; 2005 Oct 22 26; Melbourne, Australia 2005;74.

(124) Lumbreras B, Porta M, Marquez S, Pollan M, Parker LA, Hernandez-Aguado I et al. QUADOMICS: an adaptation of the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) for the evaluation of the methodological quality of studies on the diagnostic accuracy of '-omics'-based technologies. Clin Biochem 2008; 41(16-17):1316-1325.

(125) Evaluation of the methodological quality of diagnostic studies: experiences with QUADAS and suggestions for amendments. Methods for Evaluating Medical Tests and Biomarkers.Symposium; 2010.

(126) Meads CA, Davenport CF. Quality assessment of diagnostic before-after studies: development of methodology in the context of a systematic review. BMC Medical Research Methodology 2009; 9:3.

(127) Mann R, Hewitt CE, Gilbody SM. Assessing the quality of diagnostic studies using psychometric instruments: applying QUADAS. Social Psychiatry and Psychiatric Epidemiology 2008; 44:300-307.

(128) Hollingworth W, Medina LS, Lenkinski RE, Shibata DK, Bernal B, Zurakowski D et al. Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool. A preliminary assessment. Acad Radiol 2006; 13(7):803-810.

(129) Allen J, Annells M. A literature review of the application of the Geriatric Depression Scale, Depression Anxiety Stress Scales and Post-traumatic Stress Disorder Checklist to community nursing cohorts. Journal of Clinical Nursing 2009;949-959.

(130) Baker PA, Depuydt A, Thompson JM. Thyromental distance measurement : fingers don't rule. Anaesthesia 2009;878-882.

(131) Bours GJ, Speyer R, Lemmens J, Limburg M, deWit R. Bedside screening tests vs.video fluoroscopy or fibreoptic endoscopic evaluation of swallowing to detect dysphagia in

patients with neurological disorders:systematic review. Journal of Advanced Nursing 2009;477-493.

(132) Brenninkmeijer EE, Schram ME, Leeflang MM, Bos JD, Spuls P, I. Diagnostic criteria for atopic dermatitis: a systematic review. British Journal of Dermatology 2008; 158(4):754-765.

(133) Broekhuizen BD, Sachs AP, Oostvogels R, Hoes AW, Verheij TJ, Moons KG. The diagnostic value of history and physical examination for COPD in suspected or known cases:a systematic review. Family Practice 2009;260-268.

(134) Bruening W, Fontanarosa J, Tipton K, Treadwell JR, Launders J, Schoelles K. Systematic review: comparative effectiveness of core-needle and open surgical biopsy to diagnose breast lesions. Annals of Internal Medicine 2009;1-21.

(135) Bruyninckx R, Aertgeerts B, Bruyninckx P, Buntinx F. Signs and symptoms in diagnosing acute myocardial infarction and acute coronary syndrome: a diagnostic meta-analysis. British Journal of General Practice 2008; 58(547):105-111.

(136) Burr JM, Mowatt G, Hernandez R, Siddiqui MA, Cook J, Lourenco T et al. The clinical effectiveness and cost-effectiveness of screening for open angle glaucoma: a systematic review and economic evaluation. Health Technology Assessment 2007; 11(41):1-190.

(137) Cahill RA, Leroy J, Marescaux J. Could lymphatic mapping and sentinel node biopsy provide oncological providence for local resectional techniques for colon cancer: a review of the literature. BMC Surgery 2008; 8:17.

(138) Calvert E, Chambers GK, Regan W, Hawkins RH, Leith JM. Special physical examination tests for superior labrum anterior posterior shoulder tears are clinically limited and invalid:a diagnostic systematic review. Journal of Clinical Epidemiology 2009;558-563.

(139) Chan BK, Melnikow J, Slee CA, Arellanes R, Sawaya GF. Post treatment human papilloma virus testing for current cervical intra epithelial neoplasia:a systematic review. American Journal of Obstetrics and Gynecology 2009;422.

(140) Chou R, Fanciullo GJ, Fine PG, Miaskowski C, Passik SD, Portenoy RK. Opioids for chronic non cancer pain : prediction and identification of aberrant drug-related behaviors : a view of the evidence for an American pain society and American academy of pain medicine clinical practice guideline. Journal of Pain 2009;131-146.

(141) Cnossen JS, Vollebregt KC, de VN, ter RG, Mol BW, Franx A et al. Accuracy of mean arterial pressure and blood pressure measurements in predicting pre-eclampsia: systematic review and meta-analysis. BMJ 2008; 336:1117.

(142) Datta S, Everett CR, Trescot AM, Schultz DM, Adlaka R, Abdi S et al. An updated systematic review of the diagnostic utility of selective nerve root blocks. Pain Physician 2007; 10(1):113-128.

(143) Dowling S, Spooner CH, Liang Y, Dryden DM, Friesen C, Klassen TP et al. Accuracy of Ottawa Ankle Rules to exclude fractures of the ankle and midfoot in children: a meta-analysis. Acad Emerg Med 2009; 16(4):277-287.

(144) Feder G, Ramsay J, Dunne D, Rose M, Arsene C, Norman R et al. How far does screening women for domestic (partner) violence in different health-care settings meet criteria for a screening programme? Systematic reviews of nine UK National Screening Committee critieria. Health,Technology,Assessment 2009;1-113.

(145) Gibson J, McKenzie M, Shakespeare J, Price J, Gray R. A systematic review of studies validating the Edinburgh Postnatal Depression Scale in antepartum and postpartum women. Acta Psychiatrica Scandinavica 2009;350-364.

(146) Gu P, Zhao YZ, Jiang LY, Zhang W, Xin Y, Han BH. Endobronchial ultrasound-guided transbronchial needle aspiration for staging of lung cancer : a systematic review and meta-analysis. European Journal of Cancer 2009;1389-1396.

(147) Hall S, Lewith. A review of the literature in applied and specialised kinesiology. Forschende Komplementarmedizin 2008; 15(1):40-46.

(148) Henschke N, Maher CG, Refshauge KM. A systematic review identifies five "red flags" to screen for vertebral fracture in patients with low back pain. Journal of Clinical Epidemiology 2008; 61(2):110-118.

(149) Hess EP, Thiruganasambandamoorthy V, Wells GA, Erwin P, Jaffe AS, Hollander JE et al. Diagnostic accuracy of clinical prediction rules to exclude acute coronary syndrome in the emergency department setting: a systematic review. Canadian Journal of Emergency Medicine 2008; 10(4):373-382.

(150) Jing JY, Huang TC, Cui W, Xu F, Shen HH. Should FEV1/FEV6 replace FEV1/FVC ratio to detect airway obstruction:a meta-analysis. Chest 2009;991-998.

(151) Jiyong J, Tiancha H, Wei C, Huahao S. Diagnostic value of the soluble triggering receptor expressed on myeloidcells-1 in bacterial infection : a meta-analysis. Intensive Care Medicine 2009;587-595.

(152) Kelly AM, Dwamena B, Cronin P, Carlos RC. Breast cancer :sentinel node identification and classification after neo adjuvant chemotherapy - systematic review and meta analysis. Acad Radiol 2009;551-563.

(153) Koh K, List T, Petersson A, Rohlin M. Relationship between clinical and magnetic resonance imaging diagnoses and findings in degenerative and inflammatory temporomandibular joint diseases: a systematic literature review. Journal of Orofacial Pain 2009; 23(2):123-129.

(154) Kwee RM, Kwee TC. Imaging in assessing lymphnode status in gastric cancer. Gastric Cancer 2009;6-22.

(155) Leal J, Laupland KB. Validity of electronic surveillance systems : a systematic review. Journal of Hospital Infection 2008;220-229.

(156) Liang QL, Shi HZ, Wang K, Qin SM, Qin XJ. Diagnostic accuracy of adenosine deaminase in tuberculous pleurisy: a meta-analysis. Respiratory Medicine 2008; 102(5):744-754.

(157) Ling D, Zwerling AA, Pai M. GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. European Respiratory Journal 2008; 32(5):1165-1174.

(158) Maheshwari A, Gibreel A, Bhattacharya S, Johnson NP. Dynamic tests of ovarian reserve : a systematic review of diagnostic accuracy. Reproductive Biomedicine Online 2009;717-734.

(159) Mant J, Doust J, Roalfe A, Barton P, Cowie MR, Glasziou P et al. Systematic review and individual patient data meta-analysis of diagnosis of heart failure, with modelling of implications of different diagnostic strategies in primary care. Health Technology Assessment 2009;1-232.

(160) Menke J. Diagnostic accuracy of contrast-enhanced MR angiography in severe carotid stenosis : meta-analysis with meta regression of different techniques. European Radiology 2009;2204-2216.

(161) Met R, Bipat S, Legemate DA, Reekers JA, Koelemay MJ. Diagnostic performance of computed tomography angiography in peripheral arterial disease: a systematic review and meta-analysis. JAMA 2009; 301(4):415-424.

(162) Mirkhil S, Kent PM. The diagnostic accuracy of brief screening questions for psycho social risk factors of poor outcome from an episode of pain :a systematic review. Clinical Journal of Pain 2009;340-348.

(163) Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care : a meta-analysis. Lancet 2009;609-619.

(164) Nourbakhsh A, Grady JJ, Garges KJ. Percutaneous spine biopsy :a meta-analysis. Journal of Bone and Joint Surgery American volume 2008;1722-1725.

(165) Ochoa ME, del Carmen Marin M, Frutos VF, Gordo F, Latour PJ, Calvo E et al. Cuff-leak test for the diagnosis of upper airway obstruction in adults :a systematic review and meta-analysis. Intensive Care Medicine 2009;1171-1179.

(166) Puli SR, Bechtold ML, Reddy JB, Choudhary A, Antillon MR, Brugge WR. How good is endoscopic ultrasound in differentiating various T stages of rectal cancer: meta-analysis and systematic review. Annals of Surgical Oncology 2009; 16(2):254-265.

(167) Puli SR, Reddy JB, Bechtold ML, Choudhary A, Antillon MR, Brugge WR. Accuracy of endoscopic ultrasound to diagnose nodal invasion by rectal cancers :a meta-analysis and systematic review. Annals of Surgical Oncology 2009;1255-1265.

(168) Rabin RF, Jennings JM, Campbell JC, Bair MMH. Intimate partner violence screening tools :a systematic review. American Journal of Preventive Medicine 2009;439-445.

(169) Rud B, Hilden J, Hyldstrup L, Hrobjartsson A. Performance of the Osteoporosis Self-Assessment Tool in ruling out low bone mineral density in postmenopausal women: a systematic review. Osteoporosis International 2007; 18(9):1177-1187.

(170) Sutton M, Grimmer-Somers K, Jeffries L. Screening tools to identify hospitalised elderly patients at risk of functional decline: a systematic review. International Journal of Clinical Practice 2008; 62(12):1900-1909.

(171) Szadek KM, van der Wurff P, van Tulder MW, Zuurmond WW, Perez RS. Diagnostic validity of criteria for sacroiliac joint pain: a systematic review. J Pain 2009; 10(4):354-368.

(172) Tan E, Gouvas N, Nicholls RJ, Ziprin P, Xynos E, Tekkis PP. Diagnostic precision of carcinoembryonic antigen in the detection of recurrence of colorectal cancer. Surgical Oncology 2009; 18(1):15-24.

(173) Umbehr M, Bachmann LM, Held U, Kessler TM, Sulser T, Weishaupt D et al. Combined magnetic resonance imaging and magnetic resonance spectroscopy imaging in the diagnosis of prostate cancer : a systematic review and meta-analysis. European Urology 2009;575-591.

(174) van den Broek FJ, Reitsma JB, Curvers WL, Fockens P, Dekker E. Systematic review of narrow-band imaging for the detection and differentiation of neoplastic and nonneoplastic lesions in the colon. Gastrointestinal Endoscopy 2009; 69(1):124-135.

(175) Virgili Gianni AU: Menchini Francesca AU: Murro Vittoria AU: Peluso Emanuela AU: Rosa Francesca AU: Casazza Giovanni. Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy. Cochrane Database of Systematic Reviews: Protocols 2009; Issue 4.

(176) Whitlock EP, Lin JS, Liles E, Beil TL, Fu R. Screening for colorectal cancer: a targeted, updated systematic review for the U.S. Preventative Services Task Force. Annals of Internal Medicine 2008; 149(9):638-658.

(177) Wittkampf KA, Naeije L, Schene AH, Huyser J, van Weert HC. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. General Hospital Psychiatry 2007; 29(5):388-395.

(178) Geersing GJ, Janssen KJ, Oudega R, Bax L, Hoes AW, Reitsma JB et al. Excluding venous thromboembolism using point of care D-dimer tests in outpatients: a diagnostic meta-analysis. BMJ 2009; 339:b2990.

(179) Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003; 138(1):40-44.

(180) Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001; 20(3 Suppl):21-35.

Appendix 3.1: General Details of Included Reviews

Review details	Topic	Inclusion criteria defined?						Applicability			Quality Assessment		
		P	I	T	R	O	S	Index test role defined?	Restricted to studies of this role?	Inclusion restricted to patients in whom test will be used in practice?	QA conducted?	If no QA, was quality discussed?	Was QUADAS used?
Akcil et al (2008)(9)	Histology	Yes	Yes	Yes	Yes	Yes	No	Unclear	na	Yes	No	Yes	na
Allen & Annells (2009)(129)	Questionnaire	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	na	na
Baker et al. (2009)(130)	Clinical	No	Yes	Yes	Yes	Yes	Yes	Unclear	na	Unclear	No	No	na
Banal et al. (2009)(14)	Other	No	Yes	Yes	Yes	Yes	No	No	na	No	No	No	na
Bours et al. (2009)(131)	Combination	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	na	No
Brenninkmeijer (2008)(132)	Combination	Yes	Yes	Yes	Yes	Yes	Yes	No	na	No	Yes	na	Yes
Broekhuizen et al. (2009)(133)	Clinical	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes	na	Yes
Bruening et al. (2009)(134)	Histology	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	na	Yes
Bruyninckx et al. (2008)(135)	Clinical	Yes	Yes	Yes	No	Yes	No	No	na	Yes	Yes	na	Yes
Burr et al. (2007)(136)	Other	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	na	Yes
Cahill et al. (2008)(137)	Histology	Yes	Yes	Yes	No	No	No	Yes	Unclear	Unclear	Yes	na	Yes
Calvert et al. (2009)(138)	Clinical	No	Yes	Yes	No	No	No	Unclear	na	Unclear	Yes	na	No
Chan et al. (2009)(139)	Biochemical	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	na	na
Chou et al. (2009)(140)	Other	Yes	Yes	Yes	No	No	Yes	No	na	Unclear	Yes	na	No
Cnossen et al. (2008)(141)	Other	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Unclear	Yes	na	Yes
Datta et al. (2007)(142)	Other	Yes	Yes	No	No	Yes	Yes	No	na	Yes	Yes	na	Yes

Review details	**Topic**	**Inclusion criteria defined?**						**Applicability**			**Quality Assessment**		
		P	**I**	**T**	**R**	**O**	**S**	**Index test role defined?**	**Restricted to studies of this role?**	**Inclusion restricted to patients in whom test will be used in practice?**	**QA conducted?**	**If no QA, was quality discussed?**	**Was QUADAS used?**
Dowling et al. (2009)(143)	Clinical	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	na	Yes
Feder et al. (2009)(144)	Questionnaire	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	na	Yes
Geersing et al. (2009)(127)	Biochemical	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	Yes	na	Yes
Gibson et al. (2009)(145)	Questionnaire	Yes	Yes	Yes	Yes	Yes	No	Yes	Unclear	Yes	Yes	na	No
Gu et al. (2009)(146)	Combination	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	na	Yes
Hall (2008)(147)	Combination	Yes	Yes	No	No	No	Yes	No	na	No	Yes	na	Yes
Henschke et al. (2008)(148)	Clinical	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	na	Yes
Hess et al. (2008)(149)	Clinical	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	na	Yes
Jing et al. (2009)(150)	Other	No	Yes	Yes	Yes	Yes	No	No	na	No	Yes	na	Yes
Jiyong et al. (2009)(151)	Biochemical	No	Yes	Yes	No	Yes	No	No	na	No	Yes	na	Yes
Kelly et al. (2009)(152)	Histology	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	na	No
Koh et al. (2009)(153)	Combination	No	Yes	Yes	Yes	Yes	No	No	na	No	Yes	na	Yes
Kwee et al. (2009)(154)	Imaging	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	na	No
Leal et al. (2008)(155)	Combination	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Unclear	No	Yes	na
Liang et al. (2008)(156)	Biochemical	No	Yes	Yes	No	Yes	Yes	No	na	No	Yes	na	Yes
Ling (2008)(157)	Biochemical	No	Yes	Yes	Yes	Yes	Yes	No	na	No	Yes	na	Yes
Maheshwari et al. (2009)(158)	Biochemical	No	Yes	Yes	Yes	No	No	No	na	No	Yes	na	No
Mant et al. (2009)(159)	Combination across categories	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	na	Yes
Menke (2009)(160)	Imaging	Yes	Yes	Yes	Yes	Yes	Yes	No	na	Yes	Yes	na	Yes

Review details	**Topic**	**Inclusion criteria defined?**						**Applicability**			**Quality Assessment**		
		P	**I**	**T**	**R**	**O**	**S**	**Index test role defined?**	**Restricted to studies of this role?**	**Inclusion restricted to patients in whom test will be used in practice?**	**QA conducted?**	**If no QA, was quality discussed?**	**Was QUADAS used?**
Met et al. (2009)(161)	Imaging	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unclear	Yes	na	Yes
Mirkhil et al. (2009)(162)	Questionnaire	No	Yes	Yes	Yes	No	No	Unclear	na	Unclear	Yes	na	No
Mitchell et al. (2009)(163)	Clinical	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	na	No
Ngamruengphong et al. (2009)(10)	Biochemical	Yes	Yes	Yes	Yes	No	Yes	Yes	Unclear	Unclear	No	Yes	No
Nourbakhsh et al. (2008)(164)	Histology	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	No
Ochoa et al. (2009)(165)	Other	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Unclear	Yes	na	Yes
Puli et al. (2009)(166)	Imaging	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Unclear	Yes	na	Yes
Puli et al. (2009)(167)	Imaging	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	na	Yes
Rabin et al. (2009)(168)	Questionnaire	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	na	No
Rud et al. (2007)(169)	Questionnaire	Yes	Yes	Yes	Yes	No	Yes	Yes	Unclear	Yes	Yes	na	Yes
Sutton et al. (2008)(170)	Questionnaire	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	na	Yes
Szadek et al. (2009)(171)	Other	Yes	Yes	Yes	Yes	No	Yes	No	na	No	Yes	na	Yes
Tan et al. (2009)(172)	Biochemical	No	Yes	Yes	Yes	Yes	No	Yes	Unclear	No	Yes	na	Yes
Tandon et al. (2008)(11)	Histology	Yes	Yes	No	Yes	Yes	Yes	No	na	No	No	Yes	na
Umbehr et al. (2009)(173)	Imaging	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes	na	No
van den Broek et al. (2009)(174)	Imaging	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	na	Yes
Virgili et al. (2009)(175)	Imaging	Yes	Yes	Yes	Yes	No	Yes	Yes	Unclear	Yes	Yes	na	Yes

Review details	**Topic**	**Inclusion criteria defined?**						**Applicability**			**Quality Assessment**		
		P	**I**	**T**	**R**	**O**	**S**	**Index test role defined?**	**Restricted to studies of this role?**	**Inclusion restricted to patients in whom test will be used in practice?**	**QA conducted?**	**If no QA, was quality discussed?**	**Was QUADAS used?**
Whitlock et al. (2008)(176)	Combination	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	na	No
Wittkampf et al. (2007)(177)	Questionnaire	No	Yes	Yes	Yes	No	No	No	na	Unclear	Yes	na	Yes

Appendix 3.2a: Details of Quality Assessment: Reviews that used QUADAS

Review Details	QUADAS Items														Modifications/Reasons for omissions (if reported)	Additional Items	Was an informal quality assessment also applied?
	1	2	3	4	5	6	7	8	9	10	11	12	13	14			
Brenninkmeijer (2008)(132)	A	A	A	A	A	A	O	A	A	A	A	O	A	A	*Item 7:* Not relevant as both index test and reference standard based on clinical assessment *Item 12:* Not relevant as was evaluating clinical criteria	None	No
Broekhuizen et al. (2009)(133)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	None	Inclusion restricted based on single reference standard
Bruening et al. (2009)(134)	M	M	O	O	A	A	O	O	O	A	A	A	O	M	*Item 1:* Was patient recruitment either consecutive or random? Was the study free from obvious spectrum bias ; Was the study prospective? *Item 2:* Were inclusion/ exclusion criteria consistently applied to all patients *Item 14:* Was a complete set of data reported for >=85% of enrolled lesions?	1. Were >=85% of patients recruited actually enrolled?; 2. Was funding for this study provided by a source that has an obvious financial interest in the findings of the study? 3. Did the study account for inter-reader differences?	Studies had to report data for at least 50% of patients enrolled to be included. Case-control studies excluded
Bruyninckx et al. (2008)(135)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	None	No
Burr et al. (2007)(136)	M	O	A	O	A	A	A	O	O	A	A	A	A	A	*Item 1:* Split into two to cover a. selection of sample from unscreened population with low prevalence of glaucoma; b. sample representative of those referred from primary care because of suspicion of glaucoma	1. Is the technology of the test still current? 2. Did the study provide a clear definition of a positive results? 3. Was the definition of a positive test determined before the study was carried out?	No
Cahill et al. (2008)(137)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	Not stated	None	No
Cnossen et al. (2008)(141)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	NA	Was any preventative Intervention administered after uterine Doppler scanning?	No
Datta et al.	A	A	A	A	A	A	A	A	A	A	A	A	A	A	NA	Also used AHRQ criteria(7)	No

Review Details	QUADAS Items														Modifications/Reasons for omissions (if reported)	Additional Items	Was an informal quality assessment also applied?
	1	2	3	4	5	6	7	8	9	10	11	12	13	14			
(2007)(142)																(study population; adequate description of test; appropriate reference standard; blinded comparison of test and reference; avoidance of verification bias)	
Dowling et al. (2009)(143)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	None	The following items were assessed as possible sources of heterogeneity: Study design, prospective data collection, consecutive recruitment.
Feder et al. (2009)(144)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	None	No
Geersing et al. (2009)(178)	A	A	A	M	A	A	A	A	A	A	A	A	A	A	Item 4: Use of cross-sectional design (fulfilled by all studies) was assessed rather than time period. No explanation of this.	None	No
Gu et al. (2009)(146)	?	?	?	?	?	?	?	?	?	?	?	?	?	?	No details on QUADAS assessment reported.	None	Data were also extracted on study design (prospective/retrospective), consecutive enrolment and whether patients were selected on the basis of a previous positive PET or CT result
Hall (2008)(147)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	Assessed reporting using STARD(179)	No
Henschke et al. (2008)(148)	A	A	O	A	A	A	A	A	A	A	A	A	A	A	Item 3 omitted as inclusion restricted based on reference standard	No	Inclusion restricted based on single reference standard
Hess et al. (2008)(149)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	No	Consecutive recruitment used as inclusion criterion
Jing et al. (2009)(150)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None reported.	Also state the used Deville criteria(8) (which they refer to as Cochrane guidelines) but no further details reported.	No
Jiyong et al.	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	Also state the used Deville	No

Review Details	QUADAS Items														Modifications/Reasons for omissions (if reported)	Additional Items	Was an informal quality assessment also applied?
	1	2	3	4	5	6	7	8	9	10	11	12	13	14			
(2009)(151)																criteria(8) (which they refer to as Cochrane guidelines). Additional items assessed were: prospective and consecutive recruitment.	
Koh et al. (2009)(153)	M	A	O	A	O	O	O	A	A	O	O	O	O	O	Item 1: Sub categories: number of patients; type of patients; Description (disease status, prevalence, severity) Item 8: Was the classification system of TMJ diagnosis described? RDC/TMD; AAOP; other with verbatim	1. Are the results of the study valid? 2. Was the setting for the image interpretation described concerning diagnostic categories and criteria for diagnoses, number of observers, prior knowledge of the results of the clinical examination? 3. Was the method for calculating the relationship described in sufficient detail and was the method adequate?	No
Liang et al. (2008)(156)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	Also used STARD(179) to assess quality	No
Ling et al. (2008)(157)	M	O	A	O	M	O	O	O	O	M	M	O	O	O	Although the authors state that QUADAS was used only the following appear to have been assessed: *Item 1:* Study design (cross-sectional vs. case control), sampling method (convenience or random sample) *Item 3:* Appropriate reference standard *Item 5:* Complete verification *Items 10 & 11:* Blinded interpretation	None	No
Mant et al. (2009)(159)	A	A	A	A	A	A	A	A	A	A	A	O	A	A	*Item 12:* Omitted as it was unclear from study reports what clinical information was	None	No

Review Details	QUADAS Items														Modifications/Reasons for omissions (if reported)	Additional Items	Was an informal quality assessment also applied?
	1	2	3	4	5	6	7	8	9	10	11	12	13	14			
															provided within the research studies and if this was similar to the information that would be available in clinical practice. *Item 13:* was omitted for studies of BNP and NT-proBNP as the tests are automated and uninterpretable or intermediate results are unlikely to occur.		
Menke (2009)(160)	A	O	A	A	A	A	A	O	O	A	A	A	A	A	Cochrane version of QUADAS used	None	Inclusion restricted based on single reference standard
Met et al. (2009)(161)	A	A	O	A	O	O	O	A	A	A	A	A	O	O	*Item 7:* Omitted as index test was not part of ref test found unclear No reason for omission of other items	Prospective design	No
Ochoa et al. (2009)(165)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	Says that criteria based on QUADAS and that 14 items were assessed but not further details	None	No
Puli et al. (2009)(166)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	None	Inclusion restricted based on single reference standard. Data also extracted on whether the study was prospective and/or consecutive
Puli et al. (2009)(167)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	None	Inclusion restricted based on single reference standard.
Rud *et al.* (2007)(169)	A	A	O	A	A	A	A	A	A	A	A	O	A	A	*Item 3:* Omitted as inclusion restricted to single reference standard *Item 12:* Omitted as cannot be operationalised in these studies *Additional item:* Were study participants adequately described?	None	Inclusion restricted to cohort/cross-sectional studies with a single reference standard; case-control studies excluded.
Sutton et al. (2008)(170)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	None	No
Szadek et al. (2009)(171)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	None	Data were extracted on prospective data collection and

Review Details	QUADAS Items														Modifications/Reasons for omissions (if reported)	Additional Items	Was an informal quality assessment also applied?
	1	2	3	4	5	6	7	8	9	10	11	12	13	14			
																	analyses were restricted to studies scoring "yes".
Tan et al. (2009)(172)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	Also used STARD(179) guidelines to assess quality	Data were extracted on whether data collection was prospective. Sample size >100 investigated as possible source of heterogeneity.
van den Broek et al. (2009)(174)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	None	Data were extracted on Inter observer and intra observer variability
Virgili et al. (2009)(175)	A	A	A	A	A	A	A	A	A	A	A	A	A	A	None	Was the study sponsored by producers of OCT devices? Were eyes or individuals the unit of analyses?	No
Wittkampf et al. (2007)(177)	A	A	A	A	A	A	A	A	A	O	A	O	A	A	*Items 10 and 12:* Scoring of the index test was fully automated and no interpretation was involved	None	No

Appendix 3.2b: Details of Quality Assessment: Reviews that did not use QUADAS

Review Details	Equivalent QUADAS items assessed														Items that map to QUADAS (Equivalent QUADAS item)	Additional Items	Was an informal quality assessment also applied?
	1. Spectrum	2. Selection	3. Ref standard	4. Time period	5. Partial verification	6. Dif verification	7. Incorporation	8. Index ex	9. Ref ex	10. Index blind to ref	11. Ref blind to	12. Clinical data	13. Uninterpretable	14. Withdrawals			
Akcil *et al.* (2008)(9)	No QA conducted																Inclusion restricted based on single reference standard. % of insufficient material and study design (prospective vs. retrospective) investigated as possible sources of heterogeneity
Allen & Annells (2009)(129)	No QA conducted																No
Baker *et al.* (2009)(130)	No QA conducted																No
Banal *et al.* (2009)(14)	No QA conducted																No
Bours *et al.* (2009)(131)	✓			✓	✓			✓		✓	✓	✓			Independent interpretation of index test and reference standard (items 10&11) Index test independent of clinical data regarding the target condition (item 12) Reference standard applied to all patients (Item 5) Time period between index test and reference standard sufficiently short (Item 4) Valid selection of study population (Item 1) Appropriate study population (Item 1) Index test described in sufficient detail to allow replication (Item 8)	Data presented in sufficient detail to allow calculation of test performance Satisfactory definition of index test and reference standard thresholds	Review restricted to "cross-sectional" design studies, appears to mean diagnostic cohort studies
Calvert *et al.* (2009)(138)	✓				✓		✓	✓		✓					Sackett criteria(12): Independent, blind comparison with the reference standard (Item 10) Patient spectrum similar to that used in practice (Item 1) Did results of index test influence decision to	Likelihood ratios reported or sufficient data to enable their calculation	No

Review Details	Equivalent QUADAS items assessed														Items that map to QUADAS (Equivalent QUADAS item)	Additional Items	Was an informal quality assessment also applied?
	1. Spectrum	2. Selection	3. Ref standard	4. Time period	5. Partial verification	6. Dif verification	7. Incorporation	8. Index ex	9. Ref ex	10. Index blind to ref	11. Ref blind to	12. Clinical data	13. Uninterpretable	14. Withdrawals			
															perform reference standard? (Item 5) Sufficient index test details for replication (Item 8)		
Chan *et al.* (2009)(139)					✓					✓	✓				Blinded assessment (Items 10 &11) Follow-up with colposcopy for all subjects? (Item 5)	Did index test influence follow-up? Serial colposcopy?	Inclusion restricted based on single reference standard. Studies with incomplete reporting of outcomes or >30% lost to follow-up were excluded. Time between colposcopy and procedure evaluated as possible source of heterogeneity
Chou *et al.* (2009)(140)	✓		✓		✓			✓			✓				Adapted from US Preventive Services Task Force(180) and empirical studies of bias and variation: Inclusion of consecutive or random clinical series of patients (Item 1) Adequate description of symptom severity, underlying condition, and duration and doses of opioids (Item 1) Adequate description of study instrument (index test) (Item 8) Appropriate criteria as reference standard (Item 3) Evaluation of outcomes or reference standard in all patients (Item 5) Evaluation of results blinded to screening instrument (Item 11)	Evaluation of test performance in population other than that used to derive instrument. Inclusion of appropriate criteria in the instrument	Inclusion restricted to prospective studies
Gibson *et al.* (2009)(145)	✓		✓		✓					✓	✓				Based on CRD Report 4 (2001)(15): Blinding of assessors (Items 10 & 11) Comparison with a reference standard (Item 3) Differential use of reference standard (Item 5) Population spectrum (including use of case-control design) (Item 1)		Studies excluded if there was a delay of >=24 hours between administration of the EPDS (index test) and reference standard.

Review Details	Equivalent QUADAS items assessed														Items that map to QUADAS (Equivalent QUADAS item)	Additional Items	Was an informal quality assessment also applied?
	1. Spectrum	2. Selection	3. Ref standard	4. Time period	5. Partial verification	6. Dif verification	7. Incorporation	8. Index ex	9. Ref ex	10. Index blind to ref	11. Ref blind to	12. Clinical data	13. Uninterpretable	14. Withdrawals			
Kelly *et al.* (2009)(152)	✓	✓	✓		✓			✓	✓						Adapted from Honest (2002) Appropriate reference test (Item 3) Description of index test (Item 8) Description of reference standard (Item 9) Adequate description of participants: ability to generalize results was determined by means of adequacy of the spectrum composition at least age distribution, disease stage, and eligibility criteria (Items 1 & 2) Appropriate reference test(s) (Item 3) Consecutive enrolment (Item 1) Prospective design (Item 1) Complete verification by reference test (Item 5) Broad population (Item 1)	Study size Point estimates and measures of variability for the primary outcome measure Whether study can be generalised Multiple investigators	No
Kwee *et al.* (2009)(154)	✓		✓	✓				✓		✓	✓			✓	Adapted items from Kelly et al.(17) and QUADAS: Prospective study? (Item 1) Adequate reference test? (Item 2) Avoidance of disease progression bias? (Item 4) Avoidance of withdrawal bias? (Item 14) Avoidance of diagnostic review bias? (Item 10) Avoidance of test review bias? (Item 11) Avoidance of spectrum bias? (Item 1) Demographic information? (Item 1) Avoidance of selection bias (consecutive or random versus else) (Item 8) Standard execution of index test? (Item 8)	Avoidance of study examination bias? Avoidance of comparator review bias? Avoidance of observer variability?	No
Leal *et al.* (2008)(155)															No QA conducted		No
Maheshwari *et al.* (2009)(158)	✓		✓		✓			✓							Study design (prospective/consecutive) (Item 1) Recruitment (Item 1) Population(Item 1) Reference standard (Item 3) Verification bias (Item 5)	Outcomes of the study	No

Review Details	**Equivalent QUADAS items assessed**														**Items that map to QUADAS (Equivalent QUADAS item)**	**Additional Items**	**Was an informal quality assessment also applied?**
	1. Spectrum	**2. Selection**	**3. Ref standard**	**4. Time period**	**5. Partial verification**	**6. Dif verification**	**7. Incorporation**	**8. Index ex**	**9. Ref ex**	**10. Index blind to ref**	**11. Ref blind to**	**12. Clinical data**	**13. Uninterpretable**	**14. Withdrawals**			
															Index test description (Item 8)		
Mirkhil *et al.* (2009)(162)	✓		✓		✓			✓		✓					CASP programme criteria(13): Was there a comparison with a reference standard? (Item 3) Did all the patients get the diagnostic test and the reference standard? (Item 5) The results of the test of interest could not have been influenced by the results of the reference standard? (Item 10) Is the disease state of the tested population clearly described? (Item 1) Were the methods for performing the test described in sufficient detail? (Item 8)	Was there a clear question for the study to address? Question on presence of bias - unclear which question and which bias	No
Mitchell *et al.* (2009)(163)										✓	✓			✓	Blinding (Item 10 & 11) Withdrawals (Item 14)	Studies were assigned a rating from I to V based on sample size, blinding , withdrawals and undefined methodological weaknesses	No
Ngamruengphong *et al.* (2009)(10)	No QA conducted																Inclusion restricted based on single reference standard, diagnostic cohort studies. Sensitivity analysis conducted on highest quality trials – those that were blinded and published as full length articles
Nourbakhsh *et al.* (2008)(164)	No QA conducted																No
Rabin *et al.* (2009)(168)	✓		✓		✓										Modification of USPSTF criteria (180): Credible reference standard performed regardless of screening test results (Items 3 & 5)	Sample size. External validity/generalizability Study description of consenting	No

Review Details	Equivalent QUADAS items assessed														Items that map to QUADAS (Equivalent QUADAS item)	Additional Items	Was an informal quality assessment also applied?
	1. Spectrum	2. Selection	3. Ref standard	4. Time period	5. Partial verification	6. Dif verification	7. Incorporation	8. Index ex	9. Ref ex	10. Index blind to ref	11. Ref blind to	12. Clinical data	13. Uninterpretable	14. Withdrawals			
															Spectrum of IPV risk for participants (Item 1)	versus nonconsenting patients (Appropriate description and conduct of statistics.	
Tandon *et al.* (2008)(11)	No QA conducted																Inclusion restricted based on single reference standard
Umbehr *et al.* (2009)(173)	✓							✓		✓				✓	Based on Lijmer (1999)(18): Study design (cohort, cross-sectional, case-control) (Item 1) Prospective/retrospective(Item 1) Consecutive enrolment(Item 1) Index test description (Item 8) Patient spectrum (Item 1) Blinding of index test interpreters (Item 10) Withdrawals (Item 14)	Experience of index test interpreters	No
Whitlock *et al.* (2008)(176)	✓		✓		✓		✓	✓			✓		✓		USPSTF criteria supplemented by QUADAS(180): Screening test adequately described (Item 8) Credible reference standard used, performed regardless of test results (Items 3 & 5) Reference standard interpreted independently of screening test (Item 11) Indeterminate result handled in a reasonable manner (Item 13) Adequate spectrum of patients included in study (Item 1)	Screening test relevant, available for primary care Adequate sample size Administration of reliable screening test	Case-control studies, studies that used an inadequate reference standard (not defined) and those that incompletely applied the reference standard were excluded.

Appendix 3.3: Details on how quality assessment was incorporated into the review

Review details	Did the review produce scoring guidelines for at least one QUADAS item?	How were items scored?	Did the review use summary scores?	Did the review group studies according to quality?	How were the results of the QA reported?				How were the results of the QA incorporated?							
					Narrative	Table	Figure	Not reported	Inclusion restricted	Restricted analysis based on quality	Subgroup/ Sensitivity analysis	Summary in results	Meta-regression	Weight meta-analysis	Recommendation for research	Not incorporated
Reviews that used QUADAS																
Brenninkmeijer(2008)(132)	Yes	Yes/No/ Unclear	No	No	×	✓	×	×	×	×	×	✓	×	×	✓	×
Broekhuizen et al. 2009)(133)	Not stated	Yes/No/ Unclear	No	No	✓	✓	×	×	×	×	×	×	×	×	×	×
Bruening et al. (2009)(134)	Yes	Yes/no/not reported	Yes	Grouped as high, moderate, low or very low quality based on summary scores.	×	×	✓	×	×	×	×	×	×	×	×	✓
Bruyninckx et al. 2008)(135)	Not stated	Yes/No/ Unclear	No	No	✓	✓	×	×	✓	×	×	×	×	×	×	×
Burr et al. 2007)(136)	Yes	Yes/No/ Unclear	No	High quality studies: scored defined 'yes' for 5 key items	✓	✓	✓	×	×	×	✓	×	×	×	×	×
Cahill et al. 2008)(137)	Not stated	Unclear	No	No	×	×	×	×	✓	×	×	×	×	×	×	×
Cnossen et al. (2008)(141)	Yes	Yes/No/ Unclear/not applicable	No	High quality study: scored ‘yes’ on at least 4/6 key items	✓	×	✓	×	×	×	×	✓	×	×	×	×
Datta et al. (2007)(142)	Not stated	Yes/No/ Unclear	No	No	×	×	×	✓	×	×	×	×	×	×	×	✓
Dowling,S. et al. (2009)(143)	No	Yes/No/ Unclear	No	No	✓	✓	×	×	×	×	✓	×	×	×	×	×
Feder et al. (2009)(144)	No	Yes/No/ Unclear	No	Low quality: failed 3 or more QUADAS items	✓	✓	×	×	×	×	✓	✓	×	×	×	×
Geersing et al. (2009)(178)	Not stated	Yes/No/ Unclear	No	No	✓	×	✓	×	×	×	×	×	✓	×	×	×
Gu et al. (2009)(146)	Not stated	Unclear	No	No	✓	×	×	×	×	×	×	✓	×	×	✓	×

Review details	Did the review produce scoring guidelines for at least one QUADAS item?	How were items scored?	Did the review use summary scores?	Did the review group studies according to quality?	How were the results of the QA reported?				How were the results of the QA incorporated?							
					Narrative	Table	Figure	Not reported	Inclusion restricted	Restricted analysis based on quality	Subgroup/ Sensitivity analysis	Summary in results	Meta-regression	Weight meta-analysis	Recommendation for research	Not incorporated
Hall (2008)(147)	Not stated	Yes/No/ Unclear	Yes	No	✓	×	×	×	×	×	×	×	×	×	×	×
Henschke et al. (2008)(148)	Not stated	Yes/No/ Unclear	No	No	✓	✓	×	×	×	×	×	×	×	×	✓	✓
Hess et al. (2008)(149)	Not stated	Yes/No/ Unclear	Yes	No	✓	✓	×	×	✓	×	×	✓	×	×	×	×
Jing et al. (2009)(150)	Not stated	Unclear	Yes	High quality: QUADAS score of 10 or more	×	×	×	×	×	×	×	×	✓	×	×	×
Jiyong et al. (2009)(151)	Not stated	Unclear	Yes	High quality: QUADAS score of 10 or more	✓	×	×	×	×	×	×	✓	✓	×	×	×
Koh et al. (2009)(153)	Not stated	Yes/ no/ can't tell or descriptive	No	No	✓	×	×	×	×	×	×	×	×	×	✓	×
Liang et al. (2008)(156)	Not stated	Yes/No/ Unclear	Yes	High quality: 11/25 for STARD or 10/14 for QUADAS	✓	×	×	×	×	×	×	×	✓	×	×	×
Ling et al. (2008)(157)	Not stated	Descriptive categories	No	No	✓	✓	×	×	×	×	×	×	×	×	×	✓
Mant et al. (2009)(159)	Yes	Yes/No/ Unclear	No	No	×	×	✓	×	×	×	×	×	×	×	×	✓
Menke (2009)(160)	Not stated	Unclear	No	No	×	×	×	✓	×	×	×	×	×	×	×	✓
Met et al. (2009)(161)	Yes	Yes/No/ Unclear	Yes	Median summary score used to split studies into high or low quality	✓	✓	×	×	×	×	✓	×	×	×	×	×
Ochoa et al. (2009)(165)	Not stated	Unclear	Yes	High quality: QUADAS score of 10 or more	✓	✓	×	×	×	×	×	×	✓	×	×	×
Puli et al. (2009)(166)	Not stated	Unclear	No	No	✓	×	×	×	×	×	×	×	×	×	×	✓
Puli et al. (2009)(167)	Not stated	Unclear	No	No	✓	×	×	×	×	×	×	×	×	×	×	✓
Rud et al. (2007)(169)	Yes	Yes/No/ Unclear	Yes	No	✓	✓	✓	×	×	×	×	×	✓	×	✓	×
Sutton et al. (2008)(170)	Not stated	Yes/No/Unclear/ not applicable	Yes	No	✓	✓	×	×	×	×	×	×	×	×	×	×

Review details	Did the review produce scoring guidelines for at least one QUADAS item?	How were items scored?	Did the review use summary scores?	Did the review group studies according to quality?	How were the results of the QA reported?				How were the results of the QA incorporated?							
					Narrative	Table	Figure	Not reported	Inclusion restricted	Restricted analysis based on quality	Subgroup/ Sensitivity analysis	Summary in results	Meta-regression	Weight meta-analysis	Recommendation for research	Not incorporated
Szadek et al. (2009)(171)	Yes	Yes/No/ Unclear	No	No	✓	✓	×	×	✓	×	×	×	×	×	×	×
Tan et al. (2009)(172)	No	Yes/No/ Unclear	Yes	High quality 18/25 for STARD and 11/14 for QUADAS	✓	×	×	×	×	×	✓	×	×	×	×	×
van den Broek et al. (2009)(174)	Not stated	Yes/No/ Unclear	No	No	✓	✓	×	×	✓	×	×	×	×	×	✓	×
Virgili et al. (2009)(175)	Yes	Yes/No/ Unclear	No	High quality: studies with appropriate spectrum considered	×	×	×	×	×	×	✓	×	✓	×	×	×
Wittkampf et al. (2007)(177)	Yes	Yes/No/ Unclear	No	No	✓	✓	×	×	×	×	×	✓	×	×	×	×
Reviews that did not use QUADAS																
Bours et al. 2009)(131)		+/ -/ +- (partially fulfilled)	No	Studies graded based on number of items fulfilled: sufficient/doubtful/insufficient.	✓	✓	×	×	✓	×	×	×	×	×	×	×
Calvert et al. 764/id}		Yes/ no/ not available	No	High quality: studies that met all 5 criteria considered (only single study)	✓	✓	×	×	×	✓	×	×	×	×	×	×
Chan et al. (2009)(139)		Yes/No/not reported	No	No	×	✓	×	×	×	✓	×	×	×	×	×	×
Chou et al. (2009)(140)		Yes/No/ Unclear	Yes	High quality: yes for at least 5/9 criteria	✓	✓	×	×	×	×	×	✓	×	×	×	×
Gibson et al. (2009)(145)		Unclear	No	Studies assigned a grading from A to D based on quality items fulfilled.	✓	×	×	×	×	×	×	×	×	×	×	×
Kelly et al. (2009)(152)		Unclear	Yes	No	✓	✓	×	×	×	×	✓	×	✓	×	×	×
Kwee et al. (2009)(154)		1 if criterion met; 0 for no/unclear	Yes	High quality: Score of at least 60% of the maximum score	✓	✓	×	×	×	×	✓	✓	×	×	×	×
Maheshwari et al. (2009)(158)		Unclear	No	Good quality: prospective, consecutive, full verification, adequate test description.	×	×	×	×	×	×	×	×	×	×	✓	×
Mirkhil et al. (2009)(162)		Yes or no	Yes	No	✓	✓	×	×	×	×	×	×	×	×	✓	×
Mitchell et al.		Unclear	Yes	Studies graded from I to IV based on items	×	✓	×	×	×	×	×	×	✓	×	×	×

Review details	Did the review produce scoring guidelines for at least one QUADAS item?	How were items scored?	Did the review use summary scores?	Did the review group studies according to quality?	How were the results of the QA reported?				How were the results of the QA incorporated?							
					Narrative	Table	Figure	Not reported	Inclusion restricted	Restricted analysis based on quality	Subgroup/ Sensitivity analysis	Summary in results	Meta-regression	Weight meta-analysis	Recommendation for research	Not incorporated
(2009)(163)				fulfilled												
Rabin et al. (2009)(168)		Unclear	Yes	Studies graded as poor, fair, good or excellent based on summary scores	×	✓	×	×	×	×	×	×	×	×	×	✓
Umbehr et al. (776/id}		Descriptive categories	No	No	✓	✓	×	×	×	×	×	×	×	×	✓	×
Whitlock et al. (2008)(176)		Unclear	No	Yes, but details not reported – only state that “poor quality studies were excluded”	×	×	×	✓	✓	×	×	×	×	×	×	×

Appendix 3.4: Data extraction form

Review details	**Scoring**
Study ID, Author, Year	Free text boxes for each section
Population Index test(s) Comparator test (where appropriate) Target condition Reference standard Outcome	Free text boxes for each section
Was the review a Cochrane review?	Yes/No
Inclusion criteria	
Were inclusion criteria defined in terms of: Population Index test(s) Comparator test (where appropriate) Target condition Reference standard Outcome Study design	Yes/No/Not applicable
Was the proposed role of the index test defined?	Yes/No/Unclear
If yes, was the review restricted to studies that evaluated the test in this role?	Yes/No
Were inclusion criteria restricted to patients in whom the test will be used in practice?	Yes/No/Unclear
Quality	
Was study quality formally assessed? If yes, was this done in duplicate?	Yes/No/Unclear
If study quality was not formally assessed, were aspects of quality discussed in the review? If yes give brief details	Yes/No Free text
Were the criteria used to assess quality reported? If yes, extract name of tool and/or list items	Yes No Free text
If QUADAS was used, please indicate for each QUADAS item, whether the item was assessed, omitted or modified. If modified or omitted please give details of reasons why (if reported)	Assessed/Modified/Omitted Free text
If QUADAS was used, were any additional items added? If yes, please give details	Yes/No Free text
Were additional criteria used to assess applicably? If yes, please give details	Yes/No Free text

If the review used QUADAS, were review specific guidelines for scoring produced?	Yes - based on QUADAS background or Cochrane handbook Yes – developed criteria stated for at least one item Yes-state that guidelines produced No Not stated
Were individual items rated as yes/no/unclear?	Yes/No/Not stated
Did the review use summary quality scores?	Yes No
Did the review group studies as "high" and "low"? If yes, please give details	Yes/No Free text
Were data on inter-rater reliability reported? If yes, please extract.	Yes/No Free text
How were the results of the quality assessment reported? Summary table Summary figure Narrative description Recommendations for future research Not reported	Tick boxes
How was the quality assessment incorporated into the review? Inclusion restricted to studies fulfilling certain items Sensitivity analysis by quality item Restricted the primary analysis to studies at low risk of bias Included a summary with the interpretation of results As a variable in a meta-regression (either as overall score or individual components) To weight the meta-analysis Did not incorporate QUADAS into the meta-analysis or review conclusions Other - Extract details if other	Tick boxes Free text
If the review used QUADAS, were any items highlighted as being particularly problematic to apply? If yes, please give details	Yes No Free text
Any other comments in relation to quality not covered above	Free text

Appendix 4: QUADAS questionnaire and detailed summary of responses

1. Please provide the following details relating to your review topic:
1a. Participants
Adults
adults (general population, inpatients, outpatients, including elderly)
Adults and children
adults presenting to ambulatory care centre with main presenting complaint of a sore throat
adults with minor head injury (GCS13-15)
Adults with suspected Left Ventricular Systolic Dysfunction presenting in primary care
All suspects of active tuberculosis
Caucasians with signs and symptoms of haemochromatosis
Children & young adults with febrile neutropenia
children under 18 with urinary tract infection
Children with acute illness
children with an acute illness
Children with febrile neutropenia undergoing treatment for cancer
EDs from around the world
Elders > 60 yrs, caregivers
Emergency physicians
European Society for Paediatric Gastroenterology, Hepatology and Nutrition (ESPGHAN) Guideline task force
human patients
individuals with low bone density
Individuals with varying shoulder pathologies
kidney transplant recipients
mild stroke patients
Participants are women with a cervical cytology result of ASCUS (triage group I) or LSIL (triage group II), detected in the framework of cervical cancer screening.
Patients in HAT endemic areas
Patients of any age with type 1 or type 2 diabetes mellitus, with or without diabetic retinopathy
patients presenting in primary care with non-acute abdominal pain
patients presenting to the emergency department or urgent care setting with acute dyspnoea
patients presenting with clinical symptoms
Patients presenting with symptoms of heart failure
patients suspect of visceral leishmaniasis, healthy endemic controls, patients with other diseases
patients undergoing cardiac surgery
patients with abdominal symptoms
patients with colonic polyps
Patients with diabetic foot ulceration
Patients with high-grade glioma
Patients with peripheral arterial disease
Patients with stroke symptoms
Patients with suspected/confirmed pulmonary or extrapulmonary tuberculosis of all ages
Patients with symptoms suggestive of lower limb peripheral artery disease
patients consulting a GP with non acute lower abdominal complaints

People newly presenting with symptoms of bladder cancer or previously diagnosed with non-muscle invasive disease
people with diabetic retinopathy
Postmenopausal women
pregnant women
pregnant women
Pregnant women
Primary HPV screening
Primary school aged children
pulmonary TB patients
screening participants for colorectal cancer, people with symptoms suggestive of colorectal cancer
Subjects with known head and neck cancer for staging, for detection of secondary tumours, evaluation of treatment response, subjects with suspected recurrence. Subjects with unknown primary tumour for the detection of the primary tumour.
Suspected dementia patients
Suspected Stroke
Women diagnosed with early stage (1A2 to 2A) cervical cancer on the basis of loop biopsy
Women with lobular breast cancer

1b. Index test(s)
Alcohol Use Disorders Identification Test Consumption Questions (AUDIT-C)
anti-MCV
Any
Any colour vision test
Bedside ultrasonography
Blood markers
BMI, Weight, Skinfold thickness
BNP or NT-proBNP
Centor score/ signs and symptoms
Commercial serological antibody detection test
CRP
CRP
ct colonography
CTA
d-dimer
decision rules
DNA tests
Duplex ultrasound, magnetic resonance imaging, or computed tomography angiography, alone or in combination
Elder abuse screens
elements of patient history, physical examination, or laboratory test (tests that are easily accessible to GPs)
ELISA, IFAT, IHA, Lateral Flow, Latex agglutination
faecal calprotectin
FDG-PET(-CT)
full-ring PET, PET/CT, and other combinations of PET with conventional tests
Hain Genotype MTBDR
History and examination followed by BNP

history, physical examination, laboratory tests
Ig A Ig G: AGA, anti TG2, Endomysial antibodies, deamidated gliadin peptides, point of care tests
LAM urinary antigen
Magnetic Resonance Imaging
MRI
MRI
multiple signs and investigations for heart failure
narrow band imaging
optical coherence tomography
PCR
PCR, NASBA, LAMP (molecular amplification tests)
perfusion computed tomography, or perfusion-weighted magnetic resonance imaging of the brain
PET
Photodynamic diagnosis
point-of-care D-dimer tests
rapid diagnostic test (RDT)
Risk stratification rule
rK39 rapid diagnostic tests, Direct Agglutination test
Sentinel lymph node biopsy
serologic tests
Serum inflammatory markers (including CRP, PCT, IL6, IL8) as prognostic indicator for good/poor outcome
Several symptoms, anaemia, Faecal occult blood tests
signs and symptoms
spot protein-to-creatinine ratio
Structural neuroimaging with MRI
The index test is the B probe of the HC2 assay, which detects DNA of 13 high-risk HPV types in a cervical cell sample.
The Osteoporosis Self-assessment Tool (OST)
too many to name- all special tests of the shoulder
transoesophageal echocardiography
ultrasound, CT, MR
Umbilical artery Doppler
urinary white cell count, dipstick-leucocyte esterase, nitrite
Wound swab

1c. Comparator test(s)
not applicable/none
24hr urine collection
Alcohol Use Disorders Identification Test full version
anti-CCP
any other
APS diagnosis of abuse
composite reference standard
Computed Tomography

criteria for syndrome
CT
culture
Follow-up
Formal/Radiology suite ultrasonography/Clinical followup
histopathology
History and clinical examination followed by ECG
HPV
Mammography Ultrasound
Other commercial serological antibody detection test or sputum smear (pulmonary TB)
Other triage tests to select postmenopausal women for bone mineral density (BMD) measurement
Parasitology or serology + response to treatment
PCT, IL6, IL8
peripheral blood, microscopy, placental blood microscopy, PCR
Repetition of the cervical cytology test (conventional Pap test or liquid-based sample)
see above
SPECT, CT; MRT, Ultrasound, Chest x-ray, Endoscopy, colour-coded duplex-sonography, and combinations of those tests
strategy not incorporating DNA tests
Structural imaging with CT
White light cystoscopy
Wound tissue biopsy

1d. Target condition
1) intracranial injury and 2) need for neurosurgery
2 conditions: head and neck cancer, unknown primary cancer
active TB (pulmonary or extrapulmonary tuberculosis)
active tuberculosis
Acute ischaemic stroke - acute haemorrhagic stroke
aortic dissection
Atherosclerosis of the ascending aorta
Bacteraemia, significant bacterial infection, need for intensive care, etc
Bacteramia or documented infection
Bladder cancer
Celiac disease
CIN 2+
coeliac disease
Colon or rectum cancer
colonic neoplasia
colorectal cancer
Deep Vein Thrombosis
degree of stenosis / occlusion of artery
diabetic macular oedema

Diabetic retinopathy, or grading of diabetic retinopathy
drug-resistance TB
Elder abuse
Fetal and neonatal compromise
GABHS pharyngitis
heart failure
hereditary haemochromatosis
high-grade glioma
Human African Trypanosomiasis (HAT)
human leptospirosis
inflammatory bowel disease and colorectal cancer
irritable bowel syndrome
Ischaemic & Haemorrhagic stroke
ischaemic stroke
Left Ventricular Systolic Dysfunction
Leptospirosis
Melanoma
new epidodes of psychosis
Obesity
operable carotid stenosis
Osteoporosis
Osteoporosis
Pelvic lymph node metastases in early stage cervical cancer (1A2 to 2A)
Peripheral artery disease
placental malaria
preeclampsia
Presence of and extent of lobular breast cancer
Presence of histologically confirmed high-grade CIN or cervical cancer
Rheumatoid arthritis
Serious bacterial and bacterial infection
Serious disease
serious infections
shoulder pathologies- RC tear, impingement, instability, labral tear
systolic and diastolic heart failure
unhealthy alcohol use (alcohol dependence, misuse, risky drinking, combinations)
urinary tract infection
Vascular dementia
venous thrombo-embolism
Visceral leishmaniasis
Wound infection

1e. Reference standard
% body fat measured in variety of ways using various cut offs
1) For intracranial injury: CT or MRI. 2) For neurosurgery: follow-up 4 weeks after injury
24hr urine collection
Adult protective services diagnosis of abuse
advanced imaging (CTA)
Adverse perinatal outcome
Autopsy
Biopsy
biopsy duodenum, Marsh Criteria
biopsy or follow-up
Blood culture or clinical + microbiological confirmation
Bone mineral density as measured by dual x-ray absorptiometry
Clinical
clinical acumen
Clinical assessment + imaging follow-up
clinical criteria for determining presence of heart failure (e.g. ESC)
Clinical follow-up
colonoscopy
colonoscopy and biopsy
Colposcopy and histology of cervical tissue (punch or excision biopsy), accepting a negative colposcopy as ascertainment of absence of disease
composite reference standard (including CT-angio, V/Q scanning en ultrasonography)
Conventional angiography or findings at surgery/follow-up
culture
culture or smear for acid-fast bacilli in countries with estimated TB incidence rate ≥ 50/100,000 TB cases/year
diagnosis of heart failure
digital subtraction angiography
DSA
DXA
Echocardiography or coronary angiography
epiartic ultrasound scanning
follow up neuroimaging or PET
Follow-up
formal diagnoses
fulfilling the ACR criteria for RA
Fundus examination by fluorescein angiography, digital retinal photography, biomicroscopy or ophthalmoscopy (either at the time of colour vision screening for diagnostic detection studies or at follow-up for predictive studies).
fundus stereophotography or biomicroscopy
Histological diagnosis of colorectal cancer
Histology in many instances in combination with follow-up
Histopathological assessment of biopsied tissue
Imaging or Clinical+Imaging
MAT and/or culture

Microbiological confirmation
Microscopic Agglutination Test
Microscopy
No consensus about this but sometimes considered to be wound tissue biopsy
no organic disease explaning symptoms: extensive work-up
Parasitology
Pathologic analysis
pathology or clinical fup
Placental histology
small bowel biopsy & histology
standard hsitopathology
strategy not incorporating DNA tests (e.g. liver biopsy, iron studies)
surgery mostly but AC joint injection was also acceptable
Systematic pelvic lymphadenectomy, laparoscopic or open, followed by standard histological assessment of surgical specimen.
TB culture and/or molecular detection
throat swab
urine culture
variety of reference standards, including chest X-ray, blood culture, urine culture, CSF culture

2. In approximately how many reviews have you used QUADAS?

1 review:		43.8%	28
2-3 reviews:		35.9%	23
4-5 reviews:		18.8%	12
I can't remember:		1.6%	1

3. Have you used QUADAS in a Cochrane DTA review?

Yes, and the answers below relate to this review:		12.5%	8
Yes, but my most recent review is not a Cochrane DTA review:		4.7%	3
No:		82.8%	53

4. What stage is your review at?

Completed:		75.0%	48
Ongoing (Quality assessment completed):		18.8%	12
Ongoing (Quality assessment in progress):		6.2%	4

5. Prior to using QUADAS, have you been involved in the quality assessment of studies in a systematic review?

Yes:		70.3%	45
No:		29.7%	19

5.a. If yes, was this a diagnostic review?

Yes:		26.6%	17
No:		73.4%	47

6. Approximately how much time, on average, does it take you to complete a QUADAS assessment for each study? (*do not include time for general data extraction*)

Less than 5 minutes:		4.7%	3
Between 5 and 10 minutes:		29.7%	19
Between 10 and 30 minutes:		43.8%	28
Between 30 minutes and 1 hour:		17.2%	11
Between 1 and 2 hours:		4.7%	3
More than 2 hours:		0.0%	0

7. I find the amount of time it takes to complete QUADAS:

Acceptable (i.e.. the workload is balanced by perceived benefit):		89.1%	57
Unacceptable (i.e.. the workload does not justify the perceived benefit):		4.7%	3
I do not know / I am undecided:		6.2%	4

8. For each QUADAS item, please indicate whether the item was assessed, omitted, or modified. If the item was modified or omitted please provide brief details on the rationale for this and the wording of the modified item:

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

Assessed:		84.4%	54
Omitted:		6.2%	4
Modified:		9.4%	6

Please provide details on why the item was omitted, or why and how it was modified

"...patients who will receive perfusion imaging in practice?"
A normal population was screened rather than a patient group.
All studies had to recruit representative patients to be included.
allocated to "external validity"
An important methodological criterion we used was whether the recruitment was consecutive or not.
At our institution we assess external validity separately
Modified to more clearly express valid selection and representativeness of patients - given the target population of the individual study
Spectrum also described in detail in separate table
the high prevalence of coeliac disease in our selection (60%!) in symptomatic patients on average does not make it likely that patients had not been already tested/selected by a previous medical institution. Selection bias not reported in studies, this is certainly not a problem of quadas, but of the studies found.

The item was expanded to include a definition of a representative spectrum and in addition detail of exclusion criteria were requested
The review considered multiple settings and this was considered in context of review. The question in the quality assessment was whether it was a consecutive series of patients.
Wording changed to reflect the specific index and reference tests.

Item 2: Were selection criteria clearly described?			
Assessed:		84.4%	54
Omitted:		12.5%	8
Modified:		3.1%	2
Please provide details on why the item was omitted, or why and how it was modified			
allocated to "extern validity"			
criteria were specified			
Data were extracted on selection criteria and presented in the review - we did not assess this in terms of quality (valid selection was assessed under item 1)			
Not sure how to value if not provided. Often missing. So most critical domains would be helpful.			
Omitted because it was more concerned with the quality of reporting rather than methodological quality			
Part of inclusion criteria so would have scored 'yes'.			
see above, a lot of missing information			
This item is relevant but is not included in the revman version.			
This item was not part of the QA. However, it was described in the text.			
We considered there was potential overlap between this item and the first as an assessment of external validity. We wanted an assessment of the potential for selection bias. "Was inclusion of subjects based on the results of the index or comparator tests"			
We have had challenge in past systematic reviews achieving consensus on what we mean by selection criteria, for example symptoms, age, gender, HIV status. Sometimes this information is not stated as inclusion or exclusion criteria in Methods, but it is clearly provided in a Table. Also, this item not included in core quality criteria in RevMan.			
Were inclusion/exclusion criteria applied consistently? Were consecutive eligible patients enrolled?			
where inclusion and exclusion criteria clearly described			

Item 3: Is the reference standard likely to correctly classify the target condition?			
Assessed:		75.0%	48
Omitted:		17.2%	11
Modified:		7.8%	5
Please provide details on why the item was omitted, or why and how it was modified			
Acceptable reference standards were pre-defined and were part of the inclusion criteria - i.e. studies not using a recognised reference standard were excluded			
as part of inclusion criteria			
biopsy was by definition the reference standard, all other studies were excluded. We know that biopsy is far from perfect and may be worse than one of the index tests (EMA), but this problem was not picked up by QUADAS			
DXA is regarded the gold standard			

May need gradations of this--we had cases of acceptable reference standards (barely) vs. desirable/optimal.
multiple target conditions that the scanning was likely to find
Off course, that was the design of our review and an incorrect reference standard was an exclusion criterion
Only studies with DSA as ref standard were included.
the reference standard was specified in our inclusion criteria for the review, so we omitted this question
The selection criteria made this item redundant
The use of an adequate reference standard was a requirement for inclusion
There is not an agreed reference standard to classify obesity.
Was acceptable 'gold standard' used?
Was the clinical or radiological follow up >30 days after stroke onset?
We had to consider reference standards as stated in the primary studies as there was no agreed reference standard for the review topic
We were assessing two outcomes so we included this item twice, once for each item

Item 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?

Assessed:		76.6%	49
Omitted:		17.2%	11
Modified:		6.2%	4

Please provide details on why the item was omitted, or why and how it was modified
'Yes' for all studies because the index test is commonly collected with the reference standard although this is not specified
allocated to "bias"
Considered irrelevant in the context
Defined as <1 month
due to the nature of the tests both the reference standard and the index test would be performed at the same time.
For studies with follow-up as reference standard: was the follow-up appropriately long? Studies with systemic treatment of the tumour between index test and reference standard were judged to have an inappropriate reference standard.
Immediate for all - would have added nothing
often index test performed on stored (blood) samples some time after reference standard (using the same blood, but before storage)- needed to accommodate this
Pregnant women with reference standard assessed after birth
Review assessed prognostic value of markers assessed at presentation to hospital with febrile neutropenia (very acute disease). The marker tests and reference tests are necessarily closely related in time.
Sometimes outcome verification was not assessed immediately after the index/comparator test. This is considered as a weak point of the study. Nevertheless, clinically, an endpoint assessment for instance 2 years after the test allows picking up lesions not yet detectable by the gold standard at time 0.
Tests were conducted together.
The item was expanded to define what an acceptable time period was. In addition the actual time period was recorded as part of data extraction

This item was scored by defining the reference standard as being within 24 hours of injury.
thought not important to assessing this genetic test
we found one week to 2 months, but no one in the group wanted to decides whether one month or two months is already too long between index test, start of gluten free diet and improvement on biopsy. There are no data. Gain not a problem by QUADAS: we just feel that a lot of doubts are not picked up by QUADAS!
Wording changed to reflect the specific index and reference tests.

Item 5: Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?			
Assessed:		89.1%	57
Omitted:		6.2%	4
Modified:		4.7%	3
Please provide details on why the item was omitted, or why and how it was modified			
"...of the sample have clinical or radiological follow up?"			
All studies only presented patients who both received the index as well as the reference test, the group who received just one test is most of the time not mentioned in the studies			
as part of inclusion criteria			
Consider separating out two ideas that affect the likelihood of verification bias when whole sample does not receive the reference standard: 1)random sample vs. non-random and 2) proportion of sample verified. Consider specifying what proportion would be minimally adequate in terms of power and representativeness, and in terms of not needing further adjustment for verification bias. Also, if verification bias corrections were made, what methods are valid and how should these be described?			
different populations used for validity			
Our interpretation of partial and differential verification: Verification bias looms if the decision to perform the reference test is based on the result of the test under examination. In many diagnostic studies with an invasive reference test, most of the positive test results and only a small part of the negative test results are verified. Alternatively, negative test results are verified by a different, often less thorough, standard, for example follow-up. We will refer to these 2 forms of verification bias as partial verification bias and differential reference standard bias, respectively.			
selection criteria of the systematic review said, that all patients must have received DXA			
we demanded >90% of all patients had to have biopsy reported			
We would have excluded the study if it did not			
Wording changed to reflect the specific index and reference tests.			

Item 6: Did patients receive the same reference standard regardless of the index test result?			
Assessed:		82.8%	53
Omitted:		9.4%	6
Modified:		7.8%	5
Please provide details on why the item was omitted, or why and how it was modified			
"...same clinical or radiological follow up regardless..."			
Again this was specified in our inclusion criteria for the review.			
as part of inclusion criteria			
Difficulties of applying this to assess a genetic test (where genetic test is gold std - different populations used)			

In studies where histology and follow-up as reference standards for the subjects with respectively positive and negative results constituted the best possible reference standards this item was judged to be full-filled.
No invasive reference test and no accepted gold standard
See c.
see item 5, often unknown
selection criteria of the systematic review said, that all patients must have received DXA
This is our interpretation: This item refers to differential verification. If the choice of reference standard varied between individuals, score as ‘No’. Examples: a. case control study, cases have pulmonary TB confirmed by culture (reference standard); controls are healthy volunteers who receive index test and chest x-ray. B. case control study, cases have culture-confirmed pulmonary TB; controls are "healthy" volunteers who only receive the index test. Both of these examples have differential verification. In order to say 'yes' insist that the controls undergo sputum collection and culture for mycobacteria
This was split into 2 because it was possible that a different reference standard was applied but performance of the reference test was not related to the outcome of the index test
Used as an inclusion criterion
we hope that was the case and that study authors were honest by not excluding patients from the study that should have been in.
Wording changed to reflect the specific index and reference tests.

Item 7: Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?			
Assessed:		78.1%	50
Omitted:		17.2%	11
Modified:		4.7%	3
Please provide details on why the item was omitted, or why and how it was modified			
'Yes' for all studies because the index test is not part of the reference standard			
as part of inclusion criteria			
Considered irrelevant in the context			
different populations used			
DXA and index tests are performed by different technologies, so the index test cannot be part of DXA			
often not applicable			
See c.			
that was a prerequisite, all other studies were excluded			
The answer to this question was always going to be yes.			
the current practice index test was part of the reference standard. You could not receive CT or MRI without clinical acumen beforehand			
We omitted this in our pilot phase as we had defined our reference standard as CT or MRI which would have always been independent of decision rules or clinical characteristics. During the pilot phase, we realised that some of the key papers used a mixed reference standard of CT or follow-up (for intracranial injury), so we no longer applied the criteria that the ref standard had to be CT for all. This still seemed to leave this question redundant as follow-up was, in the cases we came across, independent of the decision rule or clinical characteristics. However, by the end of the project, and 80 odd papers later, I revisited my thinking, and realised that on the odd occasion, patients had been managed, and intracranial injury determined, according to a decision rule or management strategy that could be based on clinical characteristics that also formed the index test. So it could be argued that this item was relevant for the small number of papers that did this. This source of bias (being managed differently according to the results of the index test) is also picked up by item 6, so we felt it was covered adequately and continued with the item omitted to avoid scoring a paper negatively twice for the same issue.			

We would have excluded it if it were
When PET was part of the follow-up examinations this was only considered problematic in studies with a short follow-up (< 6 months) when it seemed likely that the patient hadn't developed other signs or symptoms.
Will always be yes as reference standard can only be performed after birth

Item 8: Was the execution of the index test described in sufficient detail to permit replication of the test?			
Assessed:		79.7%	51
Omitted:		12.5%	8
Modified:		7.8%	5
Please provide details on why the item was omitted, or why and how it was modified			
"Was the execution of the PWI acquisition and processing described..."			
allocated to "extern validity"			
as part of inclusion criteria			
At the time we used 11 item QADAS (see publication)			
Information was extracted and important details described - not assessed in terms of quality (descriptive item)			
More detail about Doppler added to assessment to allow different reviewers to consistently assess whether description was adequate			
Omitted because it was more concerned with the quality of reporting rather than methodological quality			
Part of inclusion criteria so would have scored 'yes'.			
very hard to answer, we had a lot of disagreement here among raters. Only laboratory staff could have answered these questions, not clinicians and not public health people			
We also included an evaluation here whether the test was performed adequately according to international standards.			
We do include this item in some of our systematic reviews. However, for the current review, all tests were commercial with package inserts or brochures describing the tests.			
We were assessing two outcomes so we included this item twice, once for each item			
Wording changed to reflect the specific index and reference tests.			

Item 9: Was the execution of the reference standard described in sufficient detail to permit its replication?			
Assessed:		79.7%	51
Omitted:		14.1%	9
Modified:		6.2%	4
Please provide details on why the item was omitted, or why and how it was modified			
"Was the execution of the clinical or radiological follow up described..."			
allocated to "extern validity"			
As for index test			
as part of inclusion criteria			
Each study reported many signs and symptoms, seldomly described in detail. As such, this item would not discriminate between good and less good studies.			
Information was extracted and important details described - not assessed in terms of quality (descriptive item). Quality of the reference test was assessed with item 3)			

Not usually an issue for culture and smear in these studies
Omitted because it was more concerned with the quality of reporting rather than methodological quality
Part of inclusion criteria so would have scored 'yes'.
We were assessing two outcomes so we included this item twice, once for each item
we were satisfied when biopsies were classified by Marsh 1992.
Wording changed to reflect the specific index and reference tests.

Item 10: Were the index test results interpreted without knowledge of the results of the reference standard?			
Assessed:		90.6%	58
Omitted:		6.2%	4
Modified:		3.1%	2
Please provide details on why the item was omitted, or why and how it was modified			
allocated to "extern validity", but only for not automated technologies			
because we only included prospective studies			
different populations used			
Index test objective			
We also evaluated whether the evaluation of the index text was blinded to the results of the comparator test and vice versa outside of the QUADAS questionnaire.			
We hope that yes and the authors were honest. Very often we did not find info on that.			
What to do if not clearly described, particularly if the diagnostic performance evaluation is a secondary aim of a primary trial asking a different question but being 'repurposed'.			
Will always be yes in an obstetric review where the index test is performed in pregnancy and reference standard performed on baby after birth			
Wording changed to reflect the specific index and reference tests.			

Item 11: Were the reference standard results interpreted without the knowledge of the results of the index test?			
Assessed:		96.9%	62
Omitted:		1.6%	1
Modified:		1.6%	1
Please provide details on why the item was omitted, or why and how it was modified			
allocated to "bias", but only for not automated technologies			
different populations used			
We hope that yes and the authors were honest. Very often we did not find info on that.			
What to do if not clearly described, particularly if the diagnostic performance evaluation is a secondary aim of a primary study being 'repurposed'.			
Wording changed to reflect the specific index and reference tests.			

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

Assessed:		76.6%	49
Omitted:		18.8%	12
Modified:		4.7%	3

Please provide details on why the item was omitted, or why and how it was modified
allocated to "bias", this is a very difficult item. Because it is not really clear, what is measured by it. external or internal validity? What is meant with practise? I think this is an item useful for doctors to ease decision-making, but it is not useful for systematic reviews.
Always present in these studies.
Can be very hard to determine in many write-ups.
Considered irrelevant in the context
Defined as duration of diabetes, hypertension, renal disease, HbA1c,smoking, visual acuity
Insufficient understanding of this item
n/a
N/A Assumed to be nil for screening test
not really applicable
Not relevant
Reporting of blinding of the assessors of both the index tests and the reference standards was poor: of 70 studies, 7 studies reported that the assessment of the index test was blinded for the results of the reference standard, 17 studies reported that the assessment of the reference standard was blinded for the results of the index test and 3 studies (4,3%) reported explicitly that they blinded the assessment of the index test for clinical information.
there was just no way to get this info from the studies.
We asked if there was blinding to clinical data, to emphasize internal validity over external validity
Wording changed to reflect the specific index and reference tests.

Item 13: Were uninterpretable / intermediate test results produced?

Assessed:		82.8%	53
Omitted:		12.5%	8
Modified:		4.7%	3

Please provide details on why the item was omitted, or why and how it was modified
All results were assessed in these designs.
allocated to "bias"
Design of study means 'intermediate' results are included in any analysis.
n/a
n/a
Needs more details. Indeterminate results can result in bias if removed, but inflation/deflation in test performance can also happen if included as positives or negatives. How can this be more precisely scored, given the risks of bias

for a specific topic?
not applicable
This is often difficult to score, independently of the subject of the review
We used: "Were uninterpretable/intermediate test results reported?"
Were at least 85% of patients accounted for? (So large chunks aren't being lost if they don't fit neatly into the 2*2 table)
Wording changed to reflect the specific index and reference tests.

Item 14: Were withdrawals from the study explained?			
Assessed:		89.1%	57
Omitted:		7.8%	5
Modified:		3.1%	2
Please provide details on why the item was omitted, or why and how it was modified			
again hard to answer and a lot of disagreement among raters. If people have to be put on a gluten free diet and have to improve symptoms in order to have CD that we believe that authors kept track of all patients for at least half a year. We were just not told!			
allocated to "bias"			
Details of missing values were included in data extraction table.			
Isn't it not just an issue of explanation--but also the absolute loss to follow-up that is important?			
modified to were withdrawals from study documented at all?			
n/a			
not mentioned in the studies, they only present the patients who received both tests.			
This is often not reported, especially in the anti-MCV review they used often retrospective data			
This item was judged positive if patients received all three: index test, comparator test and reference standard or withdrawals had to be explained			
Were at least 85% of patients accounted for?			

9. Was inter-rater reliability assessed?			
Yes:		15.6%	10
No:		84.4%	54
about 50% disagreement.			
absolute agreement			
between 70 and 80% agreement			
if both reviewers disagreed, discussion followed until agreement was reached (in 100% of cases, and if needed by consulting a third reviewer)			
informal kappa showed ~0.7 in my recollect (unable to find this piece of data). Minimal conferencing yielded near complete agreement in scoring			
None of the PDD studies reported data on observer variation			
Not formally calculated--just resolved.			
The κ statistic for interobserver variation in the initial quality assessment, before discussion with the third reviewer, was 0.53.			

Was not quantified until now. Is still going on.
we used kappa. generally >0.75

10. Did you read the background document that accompanies QUADAS before using QUADAS?			
Yes:		89.1%	57
No:		7.8%	5
I didn't know there was a background document:		3.1%	2

11. If yes, was the background document easy to understand?			
Yes:		87.3%	48
No:		12.7%	7

11.a. Please explain why
Differential and partial verification definitions were not easy to understand and use
In many instances, it remains vague how to exactly score an item. The document is open to a lot of interpretation and hence, when quality assessment is performed by multiple readers the scoring must be discussed in more specific details prior to application
My answer to this is question is really YES but I thought some more examples might have made it easier
Some explanations of question 2 and 3 were difficult to understand
somewhat. took a very long time for research assistants to grasp
We had a few discussions to decide the score of item 2 (selection criteria - in particular for studies that reported inclusion criteria but did not mention any exclusion criteria) and item 12 (same clinical data available when the test is used in practice). In clinical practice radiologists who scrutinize scans are aware of the clinical symptoms of the patients who undergo the imaging investigation. However, this is something to avoid in diagnostic accuracy studies because knowing the symptoms and clinical characteristics of stroke patients, for example, may greatly influence the way radiologists interpret scans!
Yes if assessing a study of a new test but no for use in assessing a genetic test

12. Did you read the Cochrane DTA Handbook quality chapter (chapter 9) before using QUADAS?			
Yes:		26.6%	17
No:		57.8%	37
I didn't know there was a handbook chapter on quality assessment for DTA reviews:		15.6%	10

13. If yes, was it helpful?			
Yes:		69.6%	16
No:		30.4%	7

13.a. Please explain why
did not know there was a handbook
Didn't add much to already published material, such as the BioMedCentral paper.
I didn't read it

It was not completely finished yet at that moment. But the finished part was helpful to me.
N.A.
no comment, your questionnaire asked to fill something in but we did not want to
Sorry ticked this by mistake

14. Did you produce scoring instructions specifically for your review?			
No - we did not use any guidelines for scoring QUADAS:		20.3%	13
No - we only referred to existing guidance (Cochrane Handbook or QUADAS background document):		28.1%	18
Yes - we adapted exiting guidance:		29.7%	19
Yes - we produced our own scoring guidelines:		21.9%	14

15. Did you use QUADAS to calculate a summary quality score?			
Yes:		20.3%	13
No:		79.7%	51

15.a. If yes, please give details on how this was done
1 point was assigned for each item marked Y, 1 point was deducted for each item marked N, and 0 points assigned for unreported/unclear items.
2 if the point was met, 0 if not and 1 when it was unclear
according to total items
add
Each "yes" answer was given 1 point; each "unclear" answer was given 1/2 point and we made a simple addition
I summarised total score for each study (horizontally) and overall performance for each QUADAS item. Therefore I did nor only find out the Study(s) with the highest score but the items that were best scored. I also noted the items that were unclear.
In a previous textbook that I wrote, we used QUADAS on hundreds of articles. Our experience there told us that studies that scored less than 10 were at high risk of bias
see Q 14
Studies graded Alphabetically A or B with specific "potential QUADAS limitations". Grade "A" indicated adequate blinding.
we counted the items that were considered sufficiently documented
We first converted the individual item answers to numeric scores by counting 1 for each Yes answer, -1 for each No, and -0.5 for each NR. For a 14-item modified scale, the raw score was normalized by adding 14, dividing by 28, and multiplying by 10.
we stated the QUADAS value for each study
yes +1, don’t know 0, no -1

16. Did you use QUADAS to stratify studies according to quality?			
Yes:		29.7%	19
No:		70.3%	45

16.a. If yes, please give details on how this was done
>12
according to total items
Analyses are ongoing. There are only 7 studies in the meta-analysis and we will probably limit the number of analyses to few items.
Eliminate studies with fatal flaws using USPSTF approach on top of QUADAS scoring.
essential criteria for study inclusion documented essential criteria for classification of high quality study documented
For each review determined which quality items were the most important and a study was graded as high or low quality on the basis of how many of these quality items it complied with
High 8.4+, Moderate 6.7-8.4, Low up to 6.7
high quality = QUADAS > 11
Is going on. As mentioned before quality issues categorised (0,1,2) are considered for meta-regression.
median of study score
not all 10 domains were discriminatory - some were all "unclear", which was not helpful for investigating differences in findings. we used domains where there was some discrimination to investigate findings
Not exactly. We listed the QUADAS "limitations" (obviously some with more than others) but we did not rank per se based on the number of limitations.
Studies were stratified according to the total QUADAS score (below or equal to 7 versus above 7)
This is still ongoing. Studies that scored 11-14 were very good quality, 7-10 (good) and 6-9 (fair)and 1-5 (inadequate/poor).
to compare studies above and below the median quality score
Using the quality summary score described above
We carried out a few subgroup analyses, focusing on specific items (verification, selection, and review bias
We made subgroups based on some QUADAS items. We explored by subgroup analyses whether scores on the following quality items explained variation in diagnostic performance: item 1 (validity of study sample), item 2 (test review bias), item 5 (validity of reference standard) and item 7 (differential verification bias). These items have been shown to result in biased estimates of diagnostic performance in empirical studies.

17. How were the results of the quality assessment reported?			
Summary table together with general study details:		n/a	24
Summary table of quality results alone:		n/a	29
Summary figure:		n/a	23
Narrative description:		n/a	37
Recommendations for future research:		n/a	21
Not reported:		n/a	0

18. How did you incorporate the QUADAS assessment into the meta-analysis and/or conclusions of your review?

We restricted inclusion to studies fulfilling certain QUADAS items:		n/a	9
We conducted sensitivity analysis by QUADAS item:		n/a	14
We restricted the primary analysis to studies at low risk of bias:		n/a	2
We included a summary with the interpretation of results:		n/a	31
We used QUADAS as a variable in a meta-regression (either as overall score or individual components):		n/a	10
We used summary QUADAS scores to weight the meta-analysis:		n/a	0
We did not incorporate QUADAS into the meta-analysis or review conclusions:		n/a	11
Other *(please specify)*:		n/a	13

Analyses are ongoing; there are only 7 studies in the meta-analysis and we will probably limit the number of analyses
Higher-scoring evidence bases influence strength of evidence ratings
in the results we sometimes referred to scores of studies on specific QUADAS items
No meta-analysis appropriate
No meta-analysis possible
No meta-analysis was appropriate for the included studies
Still in the process of performing a meta-analysis. The not all studies in QUADAS will be included in the meta-analysis. But in the discussion and conclusion we will discuss the quality and meta-analysis of given articles
Sub-group analysis with those studies deemed to be high quality as described in section 16 and as variable in meta-regression
We did not have sufficient homogenous data to conduct a meta analysis, so the quadas items could only be used normatively to highlight potential sources of bias.
we did not pool studies because clinical heterogeneity was to high
We included a summary description of the quality of included studies with the interpretation of findings.
we included QUADAS items that possibly resulted in bias for our main results as individual items in a meta-regression.
we used some QUADAS items to perform subgroup analysis

Section 6

19. Have you attended any training in the use of QUADAS?			
Workshop on quality assessment at Colloquium:		14.1%	9
Workshop on quality assessment at a symposium:		1.6%	1
Workshop training in Amsterdam:		3.1%	2
Training aimed at Cochrane Review Groups:		4.7%	3
I have not received any specific training:		65.6%	42
Other *(please specify)*:		10.9%	7
hands on training from Cochrane expert			
I attended symposia/conferences on diagnostic accuracy studies. Especially for our first HTA report we had very intensive internal methodological discussions.			
I do a lot of reading			
I have received instruction from one member from a Cochrane Review group. This was in a training for Evidence based Medicine and diagnostics.			
lecture on QUADAS by M. Leeflang and JB Reitsma during the MsSc Epidemiology course at Utrecht University			
Local training by expert on diagnostic systematic review within our Institute			
various			

20. Was an internal training session organised to ensure reviewers applied the tool consistently?			
Yes:		42.2%	27
No:		57.8%	37
20.a. If yes, please give details			
Agreed assessment of quality criteria and assessment of studies in duplicate with assessment of agreement			
All reviewers involved in meetings where specific questions were defined and a pilot data extraction and quality assessment conducted at the start of review.			
As described earlier the reviewers met regularly and encountered methodological issues and their possible impact on the outcome were discussed and the assessment of the study quality was standardized accordingly.			
Discussion of discrepancies of raters after extraction of first two studies.			
Discussion, guideline drawn up			
Explanation of QUADAS to less experienced reviewer			
Internal brief conference among the two reviewers			
Less a training session and more a discussion of differences in the scoring of the same articles by different evaluators			
Meetings were organised to discuss meaning and interpretation of items, and pilots were carried out on papers not included in the review			
pilot testing in some studies with subsequent discussion of discrepancies			

Practice with relevant DTA studies and then discussion.
Reviewers applied the tool to an included study and then met to discuss findings
see previous question
The pair of us met and agreed how we'd interpret it.
Use of QUADAS had been discussed in a previous non-Cochrane systematic review on a subsample of the same studies.
We agreed an SOP.
We defined ahead, what is to be assessed by each item.
We discussed how to apply and adjusted the manual.
We discussed how we would modify the QUADAS scale to a diagnostic yield systematic review and how we would use the modified scale in practice. We piloted it then compared results and had another discussion.
we discussed the document specifying how items should be scored. We scored a few articles that would not be included in the review in order to detect problems in the instruction
We scored three articles that were almost eligible for the review, but were excluded for some minor reason.
We use the same instruction sheets and have had in-house education
We used the tool independently on 3 papers and compared interpretation to develop 'decision rules'
we were 3 raters and rated 3 papers, then we discussed their QUADAS scores when we deviated. For the coming studies we specified the meaning of some questions.
we worked with a small group and the Cochrane expert
Yes,. We gathered all reviewers, went over scoring rules, and answered questions. we provided papers by Whiting and Chapter 9 Cochrane DTA manual.

21. Would specific training in the use of QUADAS be helpful?			
Yes:		68.8%	44
No:		31.2%	20

22. What format of training would you be MOST likely to access?			
Online training, including webinars:		56.2%	36
In-person workshop: 1/2 day:		21.9%	14
In-person workshop: full day:		4.7%	3
Cochrane training:		7.8%	5
Other *(please specify)*:		9.4%	6
I think it's most useful to include this as part of the DTA workshop			
I think online training is OK. A sort of certification could be required before reviewers assess study quality based on standardised pilot testing material and feedback			
If you are familiar with the methodology of diagnostic research the tool is easy to complete without specific training			
none			
Would depend on budget and willingness of management to buy in to this. My preference would be a workshop, as long as it needed to be, but online training would also be very useful, if it were free.			

23. Please rate QUADAS for the following on the five point scale:

23.a. Inclusion of all important items

Very good:	39.1%	25
Good:	50.0%	32
Average:	10.9%	7
Poor:	0.0%	0
Very poor:	0.0%	0

23.b. Ease of use

Very good:	21.9%	14
Good:	53.1%	34
Average:	25.0%	16
Poor:	0.0%	0
Very poor:	0.0%	0

23.c. Clarity of instructions

Very good:	25.0%	16
Good:	48.4%	31
Average:	23.4%	15
Poor:	3.1%	2
Very poor:	0.0%	0

23.d. Validity (whether QUADAS helped to distinguish between studies of different qualities)

Very good:	23.4%	15
Good:	46.9%	30
Average:	23.4%	15
Poor:	4.7%	3
Very poor:	1.6%	1

24. Please specify aspects of QUADAS that you DO like, and why:
1 to 12, these answer most of the important issues in quality assessment.
14 items can be done
all the quotes ware necessary
As a reviewer new to diagnostics, it was so useful to have a thoroughly researched tool to guide quality assessment and to make me think about sources of bias in this type of study.
Backed by research evidence on effects of bias
clarify
Clearly laid out and specific
consistency with ratings of RCT quality --> comprehensible
Covers key areas of bias in diagnostic reviews

Covers most important design features shown to influence the results of diagnostic accuracy studies
ease of use
Ease of use and comprehensiveness.
Ease of use and coverage of important aspects of study design.
Ease of use and does address most of the important contributors to bias
easily understood, straightforward to use
Easy to understand and use
easy to use
Easy to use and clear
Forces authors to assess study sample characteristics: very useful to understand on whom testing was conducted in each study (can influence test performance and finally generalizability of review results)
Gives opportunity to assess quality of studies in reliably subjective way.
Good coverage of main quality aspects.
good tool to spot weak points/possible sources of bias in studies; makes it easier to choose which studies to include in the meta analysis
Guidance provided. Explicit recognition of the potential need for modification of items.
Guide. Format of tool.
Inclusion of items on spectrum of patient (representative sample) and items on differential verification bias and incorporation bias
it covers relevant elements of diagnostic studies
it exists
It gives a nice overview of the total quality of the included studies (well, what is written in the papers, you never know what is done and what is not written of course)
It helped us to think of, what we need to assess. We had a lot of interesting discussion through QUADAS, which helped.
It is a generic tool to allow interstudy comparison.
It is a short checklist and can be completed quite quickly
It is a standardized way to compare studies as well as reviews. If review authors report QUADAS results by study and item, specific sources of potential bias can be easily identified from a review. We also use QUADAS as a reporting guide to supplement STARD when publishing our own diagnostic studies, and as an aid to study design to avoid biases.
It is easy to use and covers all relevant items
It provides a clear structured overview of quality aspects of studies
It provides an easy manner to qualify studies. It provides a measure of something that is very abstract It allows quick exclusion of really bad studies
It's a good starting point.
It's not overly complicated and once you have used it a few times it can be applied relatively quickly to each study
It's simple and quick
Mostly good - problems arise applying to a particular topic area (e.g. genetic tests) and literature available (studies limited in quality and quantity).
Q 5, 10, 11, 12,13 are very straightforward and easy to answer without a great need for adaptation to individual situations.
QUADAS includes most of the important design issues in diagnostic research
QUADAS provides for interpretation for each item so you can refer in the assessment process. There is choice of

rephrasing or omitting items to suit specific needs for assessment
Quick to complete (if you've read a paper properly!)
Scoring system, yes, no, unclear. Choice of quality items
Simplicity, and it asks the right questions.
specification of key questions for test based studies (in contrast to treatment based studies)
Speed, ease of use, objectivity; most quality assessment instruments require some form of subjective judgement (often largely subjective), which seems to me to remove the point of using a formalised tool. In general QUADAS items are capable of objective definition.
Standardised exploration of study quality
The checklist as a whole provides a rigorous assessment of a diagnostic test evaluation and I think combines the best bits from other checklists.
The emphasis on blinding and external validity
The problem is mostly not QUADAS but the very unclear description in studies of what was done.
the score is not only useful to assess the quality of studies in meta-analysis but is also a good guide, together with the STARD criteria, to design good quality diagnostic studies.
The support document is very clear - apart from the sections on withdrawals and uninterpretable results
Verification bias and selection bias Crucial aspects of validity for the reviews I was involved in, helped to distinguish between studies, and actually showed differences in results between studies meeting or not meeting these items
Very clear to use
Well documented
Well specified questions with clear. Well described (a. What is meant by this item; b. Situations in which this item does not apply; c. How to score this item)
Widely accepted, items generally not controversial

25. Please specify aspects of QUADAS that you DO NOT like, and why:
a global rating is missing, although most readers would like to see something like that
addition and modification of questions are required which are specific to the topic which is being reviewed.
As a broad tool, requires adaptation for different questions.
as suggested the one before last item is often difficult to complete and the last item is often scored unknown or positively because data on this item is most of the time lacking
Can be difficult for people lacking methodological expertise.
Difficult to get the same score with different users.
Does not include items to assess quality of comparative studies
I can't answer 23d because we have no way of knowing whether QUADAS can distinguish between studies of different qualities as there is no gold standard to compare it to. But you have forced me to give an inaccurate answer because of a lack of a "no answer" button. You should have piloted this questionnaire first.
I think it would be helpful to assess risk of bias and not "only" reporting. Sometimes external and internal validity got mixed up I think.
It is difficult to score the withdrawals and uninterpretable results items.
It is not always easy to state yes, no, unclear. For instance, a good study cannot have reported all elements, then it is difficult to categorise.
It is not the tool as such. It is just that quality appraisal is so difficult.
It's adaptability is great, but I would recommend that all items are scored (unless you really know what you're doing) whilst doing the review and items omitted at writing up stage if they prove to be redundant, to avoid nasty surprises!

item 13, because in some cases one reports that there were none. That is also a positive point not mentioned in the standard version.
Items such as partial verification or masking of index and reference often scored unclear
Journal Reviewers should ask the authors to comply with the 14 Qudas items and put them into the text so they can be found. It is not a problem of QUADAS but of the missing info in the studies.
Many grey areas due to poor reporting left to reviewer to sort out. Sometimes reporting issues confused with issues of study validity, i.e. were withdrawals described? That is a reporting issue only, as could be described but insufficiently handled.
Many of the items are very interpretational, so in general it is difficult to compare the assessment of one reviewer to another.
Needs to be adapted to each review/test but this would be a criticism of any quality tool
Not always easy to understand. Training will certainly be useful.
perhaps too many items
Q1: one of the most difficult items to judge as many factors play a role such as the likelihood of prior selection/testing depending on whether patients are recruited in a university hospital or a primary care setting. Q4: Doesn't fit very well when the only possible combination of reference standard is follow-up and yet the study is neither purely prognostic nor purely diagnostic. Q12: I always feel a bit ambivalent about this question for tests where a subjective judgment is made like imaging and wonder whether in settings with unblended evaluation you shouldn't have the diagnostic accuracy of a all the tests prior to the index test and then check whether you can improve the accuracy with the addition of the index test. Otherwise you could theoretically have the situation that you have an index test with no information which you simply interpret based on the information from all the previous testing.
Q6. Did patients receive the same reference standard regardless of the index test result? This may not always be appropriate for a test requiring biopsy analysis as part of the reference standard if the index test is negative therefore no suspicious material is identified for biopsy. The reference standard for test negatives would then be something else e.g. follow-up over a period of time. Q13. Were uninterpretable/intermediate tests reported? Answering Yes to this question gives the impression that there were uninterpretable/intermediate test results when this might not have been the case. Q14. Were withdrawals from the study explained? Answering Yes to this question gives the impression that there were withdrawals even when this might not have been the case.
Question 12 is difficult to understand; as some items weigh more than others, this should be taken into account in a quality score;
Same clinical data available - difficult to know partial verification/differential verification - terminology is pompous
Scoring of some questions are open to interpretation. Not sure about QUADAS inter-rater reliability.
Several of the items seem subject to interpretation and some of our research assistants find it difficult to determine how to answer some questions.
several points are more 'reporting' items than real quality items of the study design, and you can end up scoring unclear for most of your studies, because it's just not reported. Maybe these items can be changed/rewritten or may be some omitted. (items 8, 9, 13, 14)
some items (incorporation bias, bias by presence of other relevant clinical information) are difficult to assess in many articles, however this also reflects on the quality off
Some items are difficult to score, e.g. incorporation bias, or did not seem to have much impact (whether or not usual clinical information is available)
Some items geared toward assessment of studies that are perhaps too unreliable to be reasonably included in evidence base (such as use of reference standard in all patients)
Some items request a subjective summary, for example representative spectrum of patients
Some of the questions are very much open to interpretation.
The glaring omission is that there is no item addressing the large bias associated with a case-control design
The items of the questionnaire are too prone to interpretation. The questionnaire does not take study size into account. The questionnaire does not take into account technical quality of the index test (and eventual changes

herein) There is too much weight on the reference standard, often even the optimal golden standard (usually pathology) is not as good as this questionnaire suggests, unfortunately.
The question re "are the same data available as would be in clinical practice" - I always found it difficult to answer this, the problem being that the nature of the data will differ for each review topic. Perhaps a little more help could be given in the instructions??
The unclear category, although I understand why it is included, seems not very helpful in drawing together the overall quality of the studies included in the review. In practice, you treat the unclear category similarly as the no category.
There needs to be a distinction between internal and external validity. There are many other aspects of external validity not captured by QUADAS such as consideration of those operating the tests; type of technology etc that can be captured by data extraction or alternatively by expanding QUADS to incorporate more of these aspects but distinct from issues of internal validity. I think that questions about spectrum should be distinguished from questions about internal validity.
Too many mandatory items so difficult to assess which are the most important aspects of study quality.
We had to introduce a "not applicable" response category. We had to define the "spectrum of representative patients" in Q1 very narrowly to avoid wide interrater variation.
We still struggle with consistent interpretation of the quality items. I think training is an excellent idea.
Without a good deal of adaptation and adding on items QUADAS is insufficient for valid quality assessment - particularly in comparative accuracy studies.

26. Do you have any suggestions for improving QUADAS?
Add an item addressing the large bias associated with a case-control design
Add consideration of cluster randomization Consider issues related to verification bias and statistical methods to correct Consider bias issues related to study design, i.e. case-control vs. cohort designs Consider bias issues related to dropout/attrition
An item like: "Were withdrawals from the study explained?"
As many examples as possible of how different situations are rated in the guidelines would be helpful
Consider dropping the 3 questions that relate to reporting rather than methodological quality, i.e. questions 2, 8, 9.
I don't know how one would not have to do this (above questions); perhaps collating a list of all the variations on the questions could be compiled. That is, the creation of an online database. The first time I used QUADAS I did not adapt the questions (answers to this survey are in relation to the first time use), but in subsequent reviews I did and I found it was important to provide specific details to assist in answering the questions correctly.
In the explanation please give more examples when a DTA is performed on different laboratory tests. The examples I found mainly had to do with surgery, but I would prefer examples with laboratory tests.
include items to assess quality of comparative studies
It is important to maintain the ability to modify QUADAS to suit individual reviews.
Item 1 we find often very unclear due to the 2 different aspects to 'selection'. I would suggest splitting this into 2 separate items. We find that items 12, 13 and 14 are often so poorly reported in the studies we use that we end up omitting them. It would be helpful to hear whether omission of items is appropriate and under what conditions it should be done.
Make clear that each review should be accompanied by specific guidance - related to the topic of the review Maybe identify a core set of 5-7 items that are crucial for validity and always need to be assessed. Build on existing empirical data to support the selection of these items as sources of bias / heterogeneity.
make QUADAS known to Journal reviewers so they can ask authors to feed in all the info
More emphasis on adapting QUADAS to different clinical contexts
Possible inclusion of an item on inter-operator variability/experience.
QUADAS is missing items to characterize the data collection (prospective, retrospective) and the purpose of the data collection (to assess test accuracy or for some other primary purpose). Evidence suggest that studies based

on retrospectively collected data overestimates test accuracy. In our review most studies used historical data collected prospectively, but for other purposes than assessing test accuracy. Few of the study reports mentioned issues of missing data and the number of included participants was equal to the total in 2x2 tables. We found it difficult to incorporate the problem of potentially missing data in any of the QUADAS items. Finally, in its present form QUADAS is insufficient for accuracy assessment in comparative studies where the accuracy of two or more index tests is compared against a common reference test.
Review question 12; provide clear guidelines for quality scores
see question 25. In our review we included studies that evaluated more than 1 test to a reference, so we made a division in general QUADAS items that apply to the whole study, and items that could be different for each index/comparator test.
Some recommendations on how to make a global judgement on quality. It does not have necessarily be a sum score, or applicable to all reviews, but some considerations on this issue would be very helpful.
Specify how items should be scored in various types of studies to diagnostic accuracy. Incorporate study size. Incorporate an item on technological status of the index test. Try to specify how one should treat a golden standard that is not optimal.
The wording could be slightly changed on the question about whether the whole group or a random sample of the participants received the reference standard, I find it a little confusing the way it is but can't suggest anything better! Maybe did the whole group (or a random sample) receive the ref standard?
There are 2 aspects I miss most in QUADAS: how to deal with studies which have follow-up as a reference standard. I have noticed that some reviewers simply omit question Q4 because they feel it doesn't apply. QUADAS doesn't cover studies with a comparator test yet and for this kind of studies some additional items like blinding in the evaluation of the index test for the comparator test and vice versa and whether adequate statistical methods for the comparison of two tests in the same population were used might be useful
We added in a few additional questions: 1.Hypothesis clearly defined? 2.Were the patients selected in a non biased manner? 3.Statistical tests for main outcomes adequate? 4.Were data on observer variation reported and within acceptable range?
We assessed the added value of the diagnostic test over already existing tests. It would be helpful to have items similar to the Hayden scale for prognostic studies that covers something like 'was data presented on diagnostic tests already available in practice. More and more prediction rules come available with rheumatology for risk assessment of having for example rheumatoid arthritis. An extension of QUADAS using prediction methodology would be great
We include the following item on conflict of interest: 12. Was there industry involvement in the study (industry involvement)? __Unclear ___ No ___Yes If Yes, characterize type (Select one: answers ordered from least to most industry involvement) __Donation of test materials or kits __Receipt of educational support, grants, or speaking fees __Work/financial relationship (author is an employee/consultant or owns company stock) __Involvement in design, analysis, or manuscript production
When adapting the tool, I think there is massive potential for mistakes to happen, if all the items have not been fully understood. It would be useful to have some indication of weighting of the items - which are likely to produce the most serious errors? Where is there cross-over between items (I think this can happen depending on each review and how the items are adapted and defined). Items could be grouped together in different ways. e.g.. some relate to bias inherent to study design, some relate to poor reporting. A schematic representing the direction of influence different items have would be most useful if presented alongside the scoring system, or even incorporated into the revman 5 package. I don't think we made the most of applying the results of QUADAS to our results, and this would have helped us to do this better.

27. Would you use QUADAS again?			
Yes:	▬	100.0%	64
No:		0.0%	0

28. We welcome any further comments or feedback you have about QUADAS
(Instructions Q 26 had to be completed, but I really wanted to leave it out as I didn't use formal instructions.)

Enjoyed its use, found it very straightforward and superior to Jadad. Very pleased to hear of an effort to improve its strength.
I think it is great you are continuing to improve the tool
I've enjoyed using the tool and found it by and large very helpful indeed. Thanks!
It's great - carry on the good work!
Nice work. Maybe separating external from internal validity could be helpful.
Thank you for continuing to work on this project. It is a valuable tool and much appreciated!
The quality of the biospecimen is incredibly important and information to assess this aspect in studies is often omitted. Stored samples, multicentre and mulitnational studies are especially subject to variation and issues in sample integrity. By this I mean the preanlytical variables (when it was collected, how it was processed and what the storage conditions were). The Biospecimen Reporting for Improved Study Quality (BRISQ) has elements which could be incorporated into the QUADAS or used independently (not yet published but I have a draft copy if you would like to see it. It deals a lot with tissue work though. There is also another tool called SPREC but it is more to do with coding details for biospecimens.
These comments reflect several review members perspective using this instrument. Thanks for your work on this.
We used a modified version of QUADAS for a non-standard diagnostic review and it worked so I would use it again happily for a standard diagnostic review as well.
Well done!

Appendix 5.1: Search strategies

Medline on Ovid

1 exp "Sensitivity and Specificity"/
2 False Positive Reactions/
3 false negative reactions/
4 specificit$.tw.
5 false negative.tw.
6 false positive.tw.
7 accuracy.tw.
8 predictive value$.tw.
9 likelihood ratio$.tw.
10 SROC.tw.
11 receiver operat$ curve$.tw.
12 receiver operat$ characteristic$.tw.
13 ROC.tw.
14 or/1-13
15 "bias (epidemiology)"/
16 bias.tw.
17 15 or 16
18 14 and 17
19 exp "diagnostic techniques and procedures"/
20 di.fs.
21 du.fs.
22 diagnos$.tw.
23 or/19-22
24 18 and 23
25 exp animals/ not humans/
26 24 not 25
27 (2001$ or 2002$ or 2003$ or 2004$ or 2005$ or 2006$ or 2007$ or 2008$ or 2009$ or 2010$).ed.
28 26 and 27

EMBASE on Ovid <1980 to 2010 Week 13>

1 "sensitivity and specificity"/
2 diagnostic accuracy/
3 false negative result/
4 false positive result/
5 specificity.tw.
6 false negative$.tw.
7 false positive$.tw.
8 accuracy.tw.
9 predictive value$.tw.
10 likelihood ratio$.tw.
11 SROC.tw.
12 receiver operat$ characteristic$.tw.
13 receiver operat$ curve$.tw.
14 ROC.tw.
15 receiver operating characteristic/
16 or/1-15
17 exp systematic error/
18 bias.tw.
19 17 or 18
20 16 and 19
21 exp "diagnosis, measurement and analysis"/
22 di.fs.
23 diagnos$.tw.
24 or/21-23
25 20 and 24
26 (exp animals/ or nonhuman/) not human/
27 25 not 26
28 (2001$ or 2002$ or 2003$ or 2004$ or 2005$ or 2006$ or 2007$ or 2008$ or 2009$ or 2010$).em.
29 27 and 28

BIOSIS on ISI Web of Knowledge

#17 #15 and #17 [limited to 2001 to 2010]
#16 diagnos*
# 15 #12 not #13
# 13 TS=(animal* not human*)
# 12 #10 and #11
# 11 TS=bias
# 10 #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9
# 9 TS=(receiver operat*)
# 8 TS=ROC
# 7 TS=SROC
# 6 TS="likelihood ratio*"
# 5 TS="predictive value*"
# 4 TS=accuracy
# 3 TS=specificity

# 2 TS=("false negative" or "false positive")
# 1 TS=(sensitivity same specificity)

The Cochrane Methodology Register

#1 diagnos* in All Text
#2 MeSH descriptor Bias (Epidemiology) explode all trees
#3 bias in All Text
#4 (#2 or #3)
#5 (#1 and #4)

DARE on The Cochrane Library

#1 diagnos* in Title, Abstract or Keywords
#2 MeSH descriptor Bias (Epidemiology) explode all trees
#3 bias in Title, Abstract or Keywords
#4 (#2 or #3)
#5 (#1 and #4)

Appendix 5.2: Detailed data extraction tables from original review

Study details	Methods	Bias	Evidence provided*
Aldberg(2004)(21) **Study design** Real life: review **Objective** To determine the value of overall accuracy in studies of test validity. **Type of analysis** Statistical	Issues associated with the use of overall accuracy are summarised. Also reviewed 25 studies that have used overall accuracy to summarise test performance.	Disease Prevalence	When prevalence is low, overall accuracy more closely resembles specificity; when prevalence is high, overall accuracy more closely resembles sensitivity.
Arana (1990)(79) **Study design** Real life: review **Objective** To assess the effect of diagnostic methodology on the outcome of the TRH-ST in unipolar depression. **Type of analysis** statistical	The literature was reviewed.(no further details provided), the sensitivity of the TRH-ST (thyrotropin releasing hormone stimulation test) was compared between studies that used the DSM-III and the RDC as the reference standard.	Inappropriate reference standard	The sensitivity of the TRH-ST was lower when DSM-III was used as the reference standard (34.8%) than when RDC unipolar depression was used as the reference standard (51%).
Bachmann(2009)(22) **Study design** Numeric: modelling **Objective** To demonstrate the effects of spectrum and clinical review bias using a clinical example. **Type of analysis** Statistical	Using an example of 580 patients who underwent coronary angiography and in whom the index test consisted of stress ECG, the authors investigated the effect of different population compositions on the DOR by simulating 100 hypothetical study populations with different proportions of patients with typical and atypical symptoms and calculating the DOR for each population. The effect of formally incorporating data on age, sex and symtomatology into the ECG results was also investigated.	*Demographic Features*	*Proportion of patients with atypical symptoms: The DOR initially increased as the proportion with atypical symptoms increased peaking at around 60% before decreasing again.*
		Clinical Review Bias	*ECG performance after formal incorporation of age, sex, and symptomatologyy (using a logistic regression model): increased DOR*

Study details	Methods	Bias	Evidence provided*
Barber(2006)(23) **Study design** Real life: prospective diagnostic accuracy study **Objective** To develop a simple screening question for pelvic organ prolapse (POP) and to evaluate its test characteristics in high and low prevalence populations **Type of analysis** Statistical	Data from 100 women were used to identify the question or questions that most accurately identified women with advanced pelvic organ prolapse. After identifying an accurate and reliable screening question its test characteristics were evaluated in 2 additional distinct populations: a group of 120 women presenting to a tertiary care urogynecology clinic (High prior probability of POP) and 448 women presenting to a nurse practitioner for annual gynaecologic examination (Low prior probability of POP). Patients in these 2 groups each completed the screening question and underwent a POPQ examination (ref standard)	Disease Prevalence	High pre-test probability population versus low pre-test probability population Increased sensitivity, decreased specificity
Berbaum (1988) (100) **Study design** Real life: experimental **Objective** To investigate the impact of clinical history on fracture detection with radiography	The effect of knowledge of localizing symptoms and signs in the detection of fractures was studied. Forty radiographs of the extremities were examined twice by seven radiologists; the sessions were separated by 4 months. In 26 cases, a subtle fracture was present; 14 cases were normal. In half of the cases at each session, the precise location of pain, tenderness, or swelling was provided. The observer was asked to determine if the case was normal or abnormal (provide the exact location of the fracture) and to indicate the degree of confidence in the diagnosis.	Clinical review bias	Analysis of receiver operator characteristic parameters indicates that clues regarding location of trauma facilitate detection of fractures. An improvement of 6% in the area under the ROC curve, p<0.005 was found for radiologists. The improvement is based largely on an increased true-positive rate without an increased false-positive rate, regardless of the decision criteria of the radiologist (overall willingness to "over read" or "under read"). For orthopaedic surgeons the analysis of receiver operator characteristic parameters also found that clues regarding the location of trauma facilitate detection of fractures. The area under the ROC curve showed an 11% improvement, p<0.001.
Berbaum (1989)(101) **Study design** Real life: experimental **Objective** To evaluate the influence that knowledge of localising clinical signs has on the accuracy of fracture detection by orthopaedic surgeons and radiologists. **Type of analysis** statistical	The same study as that described above was repeated with a group of orthopaedic surgeons. Results obtained by the different groups of observers were compared.	Observer variation	Statistical comparison of the two experiments showed that orthopaedic surgeons depend on clinical history much more than radiologists. This was demonstrated by a statistically significant prompting-by-speciality interaction (p<0.05).

Study details	Methods	Bias	Evidence provided*
Biesheuvel(2008)(24) **Study design** Numeric: modelling **Objective** To show the advantage of the nested case-control design for DTA studies. **Type of analysis** Statistical	Used data from a DTA cohort study of 1295 consecutive patients selected on suspicion of having deep vein thrombosis (DVT). Drew nested case-control samples from the full study population with case: control ratios of 1:1, 1:2, 1:3 and 1:4 (per ratio 100 samples were taken). Diagnostic accuracy for two tests used to detect DVT in clinical practice were estimated after correcting for sampling ratios. In the analysis of the nested case-control samples, control samples were multiplied by [1/sample fraction] corresponding to the case-control ratio (1:1 = 3.48; 1:2 = 1.74; 1:3 = 1.16; 1:4 = 0.87)..	*Distorted Selection of participants*	*Estimates from nested CC versus estimates from total cohort: no difference*
Bowler (1998)(80) **Study design** Real life: diagnostic accuracy study, retrospective **Objective** To investigate the effects of including cases with other disease affecting cognition and excluding those without necropsy in the estimation of the accuracy of necropsy for confirming Alzheimer's disease. **Type of analysis** statistical	Data were taken from the University of Western Ontario Dementia Study, a registry of dementia cases with clinical and psychometric follow up to necropsy based in a university memory disorders clinic with secondary and tertiary referrals. Data were available on 307 patients; 200 (65%) had clinically diagnosed Alzheimer's disease, 12 (4%) vascular dementia, 47 (15%) mixed dementia, and 48 (16%) had other diagnoses. One hundred and ninety two of 307 cases (63%) died and 122 of 192 fatalities (64%) had necropsies. In cases without necropsy, progressive cognitive loss was used as a marker for degenerative dementia. The outcome measures of interest were the positive predictive value of a clinical diagnosis of Alzheimer's disease allowing with and without correction for cases that were not necropsied.	Partial/ differential verification bias	The clinical diagnoses differed significantly between the population who died and those who did not. In cases without necropsy, 22% had no dementia on follow up, concentrated in early cases and men, showing considerable scope for verification bias.
Boyer(2009)(81) **Study design** Real life: review **Objective** To determine whether biases influence published estimated of the performance of diagnostic tests for carpel tunnel syndrome (CTS). **Type of analysis** Statistical	23 studies on any test for CTS were included in the review and assessed for quality using QUADAS. Meta-regression based on Moses-Littenberg was used to investigated whether any of the QUADAS items and the additional item of study design influenced estimates of sensitivity, specificity and the DOR. Only 4 QUADAS items showed appropriate dispersion of results for investigation in the analysis: spectrum bias (use of CC design), test review bias, diagnostic review bias, and differential verification bias	Distorted Selection of participants	Use of case-control design (present in 14/23 studies): increased sensitivity, specificity and DOR
		Differential verification	Differential verification bias (present 4/23 studies): no effect on accuracy estimates
		Review Bias	Test review bias (present 8/23 studies): increased sensitivity and DOR; no effect on specificity Diagnostic review bias (presented 2/23 studies) - no effect.

Study details	**Methods**	**Bias**	**Evidence provided***
Boyko (1988)(82) **Study design** Numeric: modelling **Objective** To describe the expected effects of reference standard errors on the measurement of diagnostic test sensitivity and specificity. **Type of analysis** statistical	Using formulas developed to demonstrate the expected deviations due to reference standard errors of apparent diagnostic test sensitivity and specificity, the effects of varying disease prevalence on the deviations of apparent diagnostic test sensitivity and specificity were observed.	Inappropriate reference standard	*When disease prevalence was varied from 0.01 to 0.99 the apparent diagnostic test specificity was closest to the actual value at low disease prevalence, while apparent diagnostic test sensitivity coincided with the actual value at high disease prevalence. Considerable differences existed between actual and apparent values for both sensitivity and specificity at low and high disease prevalences, even when the reference standard had close to perfect performance (96% sensitivity and specificity). The greatest deviations of the apparent diagnostic test likelihood ratios from the actual value occurred at low and high disease prevalences and came closest to the actual value at disease prevalences near 50%.*
Brealey(2007)(83) **Study design** Real life: review **Objective** To determine the effect of reference standard related bias on estimates of plain radiograph reading performance using studies conducted in clinical practice **Type of analysis** Statistical	Twenty studies evaluating any of 3 reading methods of radiography with radiography as reference standard were included. Associations between bias and reading performance using SROC regression model that produces relative DOR. The following sources of bias were assessed: reference standard, partial verification, different verification, test review, reference standard review bias.	Inappropriate reference standard	Use of less valid reference standard: consultant radiologist versus radiologists of varying seniority or Consultant/Specialist registrar radiologist (RDOR 0.5, 95%CI 0.1 to 2.5)
		Partial verification	Application of reference standard depending on observer's opinion: (RDOR 0.87; 95% CI, 0.23 to 3.30)
		Differential verification	Use of different reference standards in same study: (RDOR 0.89; 95% CI 0.23 to 3.39)
		Review Bias	Reference standard review bias: increase (RDOR,3.7;95%CI,1.6 to 8.3). Test review bias: none (RDOR,1.7;95%CI,0.6to 5.1)
Burch(2006)(26) **Study design** Real life: review **Objective** To assess the impact of including case-control studies on estimates of diagnostic accuracy **Type of analysis** Narrative	SR of accuracy of 2 distinct faecal occult blood tests (FOBT) in the detection of neoplasms, including 33 studies in total. Due to presence of large heterogeneity, no (stratified) pooling of results was attempted. Ranges of sensitivities were compared in subgroup of cohort versus case-control studies.	Distorted Selection of participants	Case-control vs. cohort study: increased sensitivity

Study details	Methods	Bias	Evidence provided*
Cagle(2009)(84) **Study design** Real life: prospective diagnostic accuracy study **Objective** To estimate the accuracy of colposcopy and visual inspection with acetic acid (VIA) while minimising the effects of reference standard misclassification bias. **Type of analysis** Statistical	2594 women invited for screening, 2005 enrolled, patients underwent VIA, hc2 and liquid based cytology. Women positive on any test (n=516) had colposcopy and digital photographs, biopsy and, whenever possible, endocervical curettage (ECC). All those receiving colposcopy also received routine ECC whenever possible whether or not colposcopy was unsatisfactory, and a directed cervical biopsy from any abnormal area. In addition, random biopsies were obtained from the four cervical quadrants where there did not appear to be any neoplastic abnormality. 1839 women were included in the analysis of whom 516 had a colposcopy, of these 504 had ECC. Accuracy of VIA was estimated using the standard gold standard of colposcopy and directed biopsy and an expanded diagnosis including ECC and 4-quadrant random biopsy.	Inappropriate reference standard	Use of expanded vs. standard colposcopy: decreased sensitivity no effect on specificity. No effects were seen on sens or spec in the valuation of LBC or hc2 with either the expanded or standard reference standard.
Cecil (1996)(85) **Study design** Real life: diagnostic accuracy study, retrospective **Objective** To determine the sensitivity, specificity, positive predictive value, and negative predictive values of stress SPECT thallium testing for the detection of coronary artery disease in a large population and to correct for work-up bias in this population **Type of analysis** statistical	From a computerised data base, reports of 4354 stress SPECT thallium studies from January 1, 1986 through December 31, 1992 were reviewed. All patients with a known history of myocardial infarction or prior coronary angiography were excluded, leaving 2688 patients. From this total, 471 patients underwent coronary angiography within 90 days following stress SPECT thallium testing. Coronary artery disease was defined as a visually assessed stenosis of a coronary artery or a major branch > 50%. Of the 2688 stress SPECT thallium studies, 1265 were normal and 1423 were abnormal. For the 471 patients who underwent catheterisation within 90 days following stress SPECT thallium testing.	Partial verification bias	The "observed" sensitivity and specificity were 98 and 14%, respectively. After correction for work-up bias using a mathematical correction method (Begg[63]), the corrected sensitivity and specificity were 82 +/- 6% and 59 +/- 2%, respectively.
Ciccone (1992)(102) **Study design** Real life: experimental **Objective** To evaluate the performance of radiologists in mammographic mass screening **Type of analysis** statistical	Seven radiologists read blindly the mammograms of 45 women (two views of each breast). The films included 12 normal, 24 benign disease and 9 cancers. The readings were repeated after 2 years.	Observer variation	Variability was higher among radiologists than between the two readings of the same radiologist, but general reproducibility was moderate. Kappa values for a positive/negative classification were 0.45 at the first and 0.44 at the second reading (inter-observer comparisons). For the intra-observer comparisons, kappa values ranged from 0.35 to 0.67. A slight increase in sensitivity was observed at the second reading. Sensitivity ranged from 33.3 - 85.7 at first reading and from 44.4 to 88.9 at second reading. Specificity ranged from 52.9 - 73.5 at first reading and from 50.0 to 80.0 at second reading.
Clark(2004)(27) **Study design**	27 DTA studies included in SR; 16 had immediate histological verification and 11 had delayed verification	Distorted Selection of participants	At least one of the following features: adequate recruitment, appropriate spectrum, or adequate blinding versus none of the above: decreased DOR

Study details	Methods	Bias	Evidence provided*
Real life: review **Objective** To empirically evaluate bias in estimation of accuracy associated with delay in verification of diagnosis among studies evaluating tests for predicting endometrial hyperplasia **Type of analysis** Statistical	by >24 hours. The effect of this delay in verification on estimates of accuracy was assessed using meta-regression based on the DOR.	Disease Progression	Delayed verification vs. immediate verification: decreased DOR
Cohen (1987)(103) **Study design** Real life: experimental **Objective** To assess the influence of training and experience on the interpretation of fine-needle aspiration biopsy (FNAB) specimens **Type of analysis** statistical	50 cases were selected from the cytology registry of the University of California, San Francisco. Each case had histologic follow-up on the course of the breast mass and the examination was assumed to provide a definitive diagnosis. 31 cases involved benign masses and 19 involved malignant masses, some cases were unusual and difficult others were straightforward. FNAB specimens from each case were examined by five observers with varying degrees of training and expertise, two were labelled as experts and the other were non-experts. ROC curves were used to investigate observer variability.	Observer variation	The ROC curves showed that training and experience significantly influenced interpretation of breast FNAB specimens. The two experts operated at a higher level of sensitivity and specificity than the three non-experts. Pairwise comparison of areas under the ROC curves showed significant differences between the experts and non-experts.
Corley (1997)(104) **Study design** Real life: experimental **Objective** To establish a histologic diagnosis of pneumonia by consensus of a panel of pathologists, to test the interobserver and intra-observer variation in the histologic diagnosis of pneumonia, to compare the diagnostic accuracy of diagnosing pneumonia with and without pre-selected histologic criteria, and to establish more specific histologic criteria for the diagnosis of pneumonia. **Type of analysis** statistical	The study group consisted of 39 patients who died after a mean of 14 days of mechanical ventilation. A post-mortem open lung biopsy was performed on all patients. The tissue was reviewed independently by four pathologists who categorised the slides from each patient as showing or not showing pneumonia. Interobserver variation was calculated using the kappa statistic. Six months following the initial evaluation, the same slides were resubmitted to one of the pathologists for re-evaluation to look for intra-observer error. Finally, the slides were reviewed and categorised by the criteria of Johanson et al into no pneumonia, mild, moderate, or severe bronchopneumonia. A comparison was made of the patients selected as demonstrating histologic pneumonia by each of the examinations.	Observer variation	The reliability coefficient (kappa) measuring agreement among the four pathologists was good at 0.916. However, the prevalence of pneumonia as determined by each of the four pathologists varied; pathologist A, 15 of 39 (38%); pathologist B, 12 of 39 (31%); pathologist C, 9 of 39 (23%); and pathologist D, 7 of 39 (18%). Resubmitting the same slides to the same pathologist 6 months later resulted in reclassification of 2 of 39 patients. Using the histologic criteria of Johanson and colleagues, 14 patients were selected as having pneumonia compared with only nine patients selected by consensus of three of four pathologists. Unanimous decisions among the observers were present in 30 patients (77%).

Study details	Methods	Bias	Evidence provided*
Cuaron (1980)(105) **Study design** Real life: experimental **Objective** To determine the possible bias of experience on the correct interpretation of Tc-99m phosphate myocardial imaging in patients with acute pericardial chest pain from diverse causes. **Type of analysis** statistical	Without prior knowledge of the significant clinical data, 6 observers independently evaluated a consecutive series of 250 myocardial scans made with Tc-99m-labeled phosphates: 127 with MDP and 23 with PPi. Of the 226 patients, all having acute pericardial chest pain, 169 were shown to have acute myocardial infarction while 57 suffered acute distress from other causes. The 6 observers, varying in their experience with nuclear medicine, compared the intensity of uptake in the heart with that in bone, and rated their impression of a 'positive' image by a 6-category scale - that is, one with 5 criterion levels. Results were expressed as receiver operating characteristic (ROC) curves, from which the optimal individual criterion level for each observer was determined.	Observer variation	The authors found very high interobserver variability in the perception of the shades of myocardial concentration, although they were based on strict and apparently objective criteria. This variability has a direct influence on the overall performance of each observer. In every instance, PPi was demonstrated to be a better tracer than MDP for myocardial imaging. The bias of the experience, visual perception, and psychology of the observer at the time of the reading of the images seems to be significant, as is the presence of uncorrected visual defects. These results justify the setting of special programs to evaluate periodically the performance of every physician who interprets studies, to establish his optimal individual criterion level instead of using a fixed criterion level to decide whether an image is 'positive'. Sensitivity in the case of PPi varied between 62-8-90% between observers and specificity varied between 79-93%.
Curtin (1997)(28) **Study design** Real life: diagnostic accuracy study, retrospective **Objective** To evaluate the accuracy of body mass index (BMI) in the diagnosis of obesity, and to investigate the presence of spectrum bias. **Type of analysis** statistical	226 Caucasians were recruited into the study. Fat, lean and bone masses were measured by dual-energy x-ray absorptiometry and BMI was calculated. The validity of the BMI for obesity was determined by its sensitivity and specificity for the whole sample and for sex and weight subgroups.	Demographic features	Overall sensitivity was 13.3% and specificity was 100%. Results for sensitivity and specificity were consistent for females and males. Overall sensitivity was equal to 0 in the subgroup weighing less than 60kg and increased up to 54.6% in the subgroup weighing more than 80kg. The major increase in sensitivity for both sexes occurred for participants weighing >=80kg. In the subgroup weighing >60 kg the sensitivity was higher in females than in males. In both sexes and in all subgroups the specificity was 100%, but the lower bound of the 95% confidence interval systematically declined in subgroups of increasing weight. The variability of sensitivity across subgroups of weight persisted when changing the cut-off for obesity. Sensitivity was higher in heavier participants than among lighter ones.
Davey(2006)(76) **Study design** Real life: review **Objective** To assess the performance of liquid-based cytology relative to conventional cytology in primary studies assessed to be of low, medium, or high methodological quality and to evaluate the effect of study design and quality on accuracy. **Type of analysis** Statistical	56 primary studies were reviewed and assessed with strict methodological criteria. Liquid-based cytology and conventional cytology were compared in terms of the percentage of slides classified as unsatisfactory, the percentage of slides classified in each cytology category, and the accuracy of detection of high-grade disease. Data were examined for studies overall and in strata to examine the effect of study quality on results. Formal analyses of the effect of quality was however not done, due to small number of trials allowing the calculation of sensitivity and specificity and large between study heterogeneity.	Test Technology	Liquid based cytology compared to conventional cytology: no effect on sensitivity, specificity or DOR

Study details	Methods	Bias	Evidence provided*
De Neef (1987)(86) **Study design** Numeric: modelling **Objective** To analyse the effect of misclassification errors on the measured accuracy of new rapid antigen detection tests for streptococcal pharyngitis. **Type of analysis** statistical	Uses models to vary the sensitivity of the reference standard from 0.9 to 1.0 and the specificity from 0.96 to 1.0. The sensitivity of the new test was varied from 0.81 to 0.95 and the specificity from 0.91 to 1.0 (the range in values reported from clinical studies). The effects of errors in the reference standard were investigated as prevalence varied.	Inappropriate reference standard	*When the new test was assumed to be more accurate than the reference standard both sensitivity and specificity were underestimated, the degree of error in the estimates was strongly related to disease prevalence.* *When the sensitivity and specificity of the new test were 95% and the sensitivity and specificity of the reference standard were increased from 96% to 98% to 100% the effects of improving the standard of comparison can be seen. The apparent sensitivity of the new test at low prevalence is much lower than the actual sensitivity. Large errors in the apparent specificity occur at high prevalence. Only in the case where the hypothetical culture is error-free are the apparent sensitivity and specificity of the new test correct (and the same for all estimates of disease prevalence).*
Detrano (1988)(29;29) (30) **Study design** Real life: review **Objective** To use meta-analysis to determine which factors affect the sensitivity and specificity of exercise thallium scintigraphy **Type of analysis** statistical	Studies involving study groups undergoing exercise thallium scintigraphy and coronary angiography performed on 50 patients or more were included in the review. Reports that did not allow calculation of sensitivity or specificity were excluded. 56 reports were included. The association of categorical variables with sensitivity and specificity was investigated using analysis of variance. Weighted linear regression of sensitivity and specificity was performed separately for each continuous variable. Stepwise weighted multiple regression was performed using sensitivity and specificity as dependent variables. Variables investigated were: % men, year of publication, angiographic definition of disease, inclusion of patients with previous MI, adequate definition of study group, avoidance of limited challenge group, avoidance of workup bias, blinding of test and reference standard, technical details.	Demographic features	Mean age and use of beta blocking medication did not affect test performance. Sex was significantly associated with sensitivity but not specificity. Percentage of men and previous MI were significantly associated with sensitivity in the multivariate analysis. Adequate definition of study group had non-significant effects on sensitivity and specificity.
		Disease severity	The percentage of patients with prior MI had the highest correlation with sensitivity, sensitivity was highest in studies that included previous MI.
		Distorted selection of participants	Avoidance of limited challenge group had non-significant effects on sensitivity and specificity.
		Inappropriate reference standard	Angiographic disease verification was not significantly related to test performance. Sensitivity and specificity were higher in studies that used tomographic imaging, only sensitivity was significantly higher. Tomographic imaging was significantly associated with sensitivity and specificity in the multivariate analysis.
		Test technology	Automation of the reading of the scintigraphic improved sensitivity but decreased specificity, differences were significant.
		Disease progression bias	The maximum interval between scintigraphy and angiography was not associated with test performance.
		Test execution	Exercise protocol was not significantly related to test performance.
		Partial verification bias	Workup bias negatively affected specificity but did not affect sensitivity.
		Review bias	Blinding of both the thallium scintigram and the coronary angiogram tended to decrease the agreement between the two, the effect of blinding was significant for sensitivity. Blinding showed a significant association with sensitivity in the multivariate analysis. For blinded studies sensitivity was 82.9% compared to 86.6% in non-blinded studies.

Study details	Methods	Bias	Evidence provided*
Detrano (1989)(31) **Study design** Real life: review **Objective** To evaluate the variability in the reported accuracy of the exercise electrocardiogram (ECG) for predicting severe coronary disease. **Type of analysis** statistical	Meta-analysis was applied to 60 consecutively published reports comparing exercise induced ST depression with coronary angiographic findings. Both technical and methodological factors were analysed. Multivariate regression analysis was used to investigate the association of technical and methodological factors with sensitivity and specificity.	Demographic features	Wide variability in sensitivity (range 40-100%) and specificity (range 17-100%) was found. Variables found to be significantly and independently related to sensitivity were: the exclusion of patients with right bundle branch block, and the exclusion of patients taking digitalis. Adjustment of exercise-induced ECG changes for changes in heart rate were strongly associated with the specificity for critical disease. Factors found not to be associated with sensitivity or specificity were: Exclusion of women, left ventricular hypertrophy, left bundle branch block and rest repolarisation abnormalities, patients taking beta-blocking agents.
		Inappropriate reference standard	The comparison with another exercise test thought to be superior in accuracy was found to be significantly and independently related to sensitivity.
		Partial verification bias	Whether the authors complied with all of the following standard: avoidance of workup bias was not associated with test performance.
		Review bias	Whether the authors complied with all of the following standard: blind reading of angiogram, blind reading of exercise ECG, was not associated with test performance.
		Handling of indeterminate results	How equivocal or non-diagnostic tests were interpreted (either excluded from analysis, included and considered as normal tests or included and arbitrary decision made as to normality) was not significantly associated with test performance.
DiMatteo(2001)(32) **Study design** Real life: retrospective diagnostic accuracy study **Objective** To assess spectrum bias of a rapid antigen tests for group A beta-haemolytic streptococcal (GABHS) pharyngitis in adults using throat culture as the reference standard. **Type of analysis** Statistical	Laboratory and clinical records from 498 consecutive adults who underwent a rapid antigen test were reviewed retrospectively. Patients were stratified according to the number of clinical features present using modified Centor criteria.	Disease Severity	Increasing Centor criteria: increased sensitivity

Study details	Methods	Bias	Evidence provided*
Diamond (1992)(88) **Study design** Numeric: modelling **Objective** To quantify the effects of various degrees of verification bias on the calculation of predictive accuracy using Bayes' theorem. **Type of analysis** statistical	A series of computer simulations was performed to quantify the effects of various degrees of verification bias on the calculation of predictive accuracy using Bayes' theorem.	Partial verification bias	*The magnitudes of the errors in absolute % differences in the observed true-positive rate (sensitivity) and false-positive rate (the complement of specificity) ranged from +11% and +23%, respectively (when the test response and the concomitant information vector were conditionally independent), to +16% and +48% (when they were conditionally non-independent). These errors produced absolute underestimations as high as 22% in positive predictive accuracy, and as high as 14% in negative predictive accuracy, when analysed by Bayes' theorem at a base rate of 50%. Mathematical correction for biased verification based on the test response using a previously published algorithm significantly reduced these errors by as much as 20%. These data indicate 1) that selection bias significantly distorts the determination of predictive accuracies calculated by Bayes' theorem, and 2) that these distortions can be significantly offset by a correction algorithm.*
Diamond (1991)(87) **Study design** Numeric: modelling **Objective** To assess the ability of the Begg-Greenes method to correct for diagnostic and prognostic selection bias, and to define the degree to which selection bias associated with the concomitant information vector affects this correction **Type of analysis** statistical	A series of computer simulations were performed to quantify the effects of various degrees of selection base on the observed true-positive rate (sensitivity), false positive rate (1-specificity) and discriminant accuracy (area under the ROC curve). Each simulation consisted of 10 000 hypothetical patients undergoing a hypothetical test with an actual true-positive rate of 80% and an actual false-positive rate of 20% with respect to an arbitrary clinical outcome. Selection bias as a result of the test response was quantified by varying the odds with respect to referral for verification from 1 to 10. Selection bias secondary to the concomitant information vector was quantified in the same way as primary selection bias, by varying the odds of referral for verification between 1 and 10. The observed true-positive and false-positive rates for the test were computed from the select subset of patients referred for verification. The discriminant accuracy of the test was assessed from the actual true and false positive rates and from the observed true and false positive rates in terms of the area under the ROC curve.	Partial verification bias	*Discriminant accuracy was assessed in terms of area under a ROC curve. Biased values of true- and false- positive rates were distributed along the curve defined by the actual true- and false-positive rates of the test for both diagnosis and prognosis. As a result, the areas under the ROC curves calculated from biased true- and false-positive rates were within 2% of the areas calculated from the actual rates. These data indicate that:* *1. Selection bias significantly distorts the determination of diagnostic and prognostic test accuracy in directionally opposite ways* *2. The distortion can be partially offset by a previously published mathematical algorithm* *3. The area under the ROC curve is insensitivity both to the primary bias associated with the test response itself and to the secondary bias associated with concomitant clinical information under a variety of circumstances.* *The direction of the bias raised estimates of sensitivity and lowered estimates of specificity.*

Study details	**Methods**	**Bias**	**Evidence provided***
Doubilet (1981)(106) **Study design** Real life: experimental **Objective** To investigate the effect of clinical information on interpretation of radiographs. **Type of analysis** statistical	Test films were included in the daily work load of readers who were unaware that a study was being carried out. Eight subtle but unambiguous abnormalities (3 lung nodules, lobar collapse, lung cyst, rib destruction, dilated oesophagus, congestive heart failure) were included on the test films. For each abnormality there were four readings with a suggestive and four with a non-suggestive clinical history. The readers were radiology residents and all interpretations were reviewed and sometimes altered by staff radiologists.	Clinical review bias	There was a statistically significant (p<0.01) increase in the rate of true-positive readings in the presence of a suggestive as compared to non-suggestive history: 16-74% for residents' readings, and 38-84% for combined resident-staff readings. There was some concomitant increase in false positives.
Egglin (1996)(33) **Study design** Real life: experimental **Objective** To determine whether radiologists' interpretations of images are biased by their context and by prevalence of disease in other recently observed cases. **Type of analysis** statistical	Methods.-A test set of 24 right pulmonary arteriograms with a 33% prevalence of pulmonary emboli (PE) was assembled and embedded in 2 larger groups of films. Group A contained 16 additional arteriograms, all showing PE involving the right lung, so that total prevalence was 60%. Group B contained 16 additional arteriograms without PE so that total prevalence was 20%. Six radiologists were randomly assigned to see either group first and then ''cross over'' to review the other group after a hiatus of at least 8 weeks. The direction of changes in a 5- point rating scale for the 2 readings of each film in the test set was compared with the sign test; mean sensitivity, specificity, and areas under receiver operating characteristic (ROC) curves were compared with the paired t test.	Disease prevalence	Results.-In the context of group A's higher disease prevalence, radiologists shifted more of their diagnoses toward higher suspicion than expected by chance (P=.03, sign test). In group A, mean sensitivity for diagnosing PE was significantly higher (75% vs. 60%; P=.04), and area under the ROC curve was significantly larger (0.88 vs. 0.82; P=.02). Conclusions.-Radiologists' diagnoses are significantly influenced by the context of interpretation, even when spectrum and verification bias are avoided. This ''context bias'' effect is unique to the evaluation of subjectively interpreted tests, and illustrates the difficulty of obtaining unbiased estimates of diagnostic accuracy for both new and existing technologies. Overall specificity was similar in both groups (64% vs. 68%).
Eldevick (1982)(107) **Study design** Real life: experimental **Objective** To assess the effect of clinical bias on the interpretation of myelography and spinal computed tomography **Type of analysis** statistical	Spinal computed tomograms and myelograms of 107 patients with sciatica or low back pain were interpreted with and without knowledge of clinical history, they were interpreted by different people on the two occasions.	Clinical review bias	90% of CT and 88% of myelographic interoperations were unchanged by knowledge of the clinical history. 11/107 CT interpretations and 12/103 myelographic interpretations differed between the first and second reading. More studies were interpreted correctly without the clinical history than with it. Knowledge of the clinical history increased the number of false-positive and decreased the number of false negative diagnoses. This study suggests a tendency of observers to interpret questionable myelographic or computed tomographic findings as positive when they correlate with clinical findings NB as the observer was different the second time round these findings could be due to interobserver variation
Elie(2008)(34) **Study design** Real life: prospective diagnostic accuracy study	1781 women had a cervical smear test (index test) and colposcopy followed by biopsy if abnormalities were detected (reference standard). Women were also	Demographic Features	Positive test for HPV (sens increased, spec decreased) and age >35 years (sens no effect, spec decreased). No association: smoking, European origin, higher educational level, menopausal status and type of contraception.

Study details	Methods	Bias	Evidence provided*
Objective To isolate factors that independently affect the accuracy of a test using an example based on the Papanicolaou smear test for detection of cervical cancer. **Type of analysis** Statistical	evaluated by the HPV test which was considered as a possible spectrum variable. Women were either attending for routine smears (screening) or were being referred for previously detected abnormality (referral). Smear tests were read twice : based on normal conditions (clinical) and reading blind to context and clinical history by two independent pathologists (optimised). Relevant patients characteristics were recorded. Sensitivity, specificity and LRs were calculated overall and stratified according to various factors. Logistic models were used to evaluate sensitivity and specificity and likelihood ratios and to identify factors independently affecting test performance.	Prior testing	Positive test for HPV: sens increased, spec decreased
		Disease Prevalence	Referral setting vs. screening: increased sensitivity, decreased specificity
		Clinical Review Bias	Clinical reading vs. optimised interpretation (blinded to clinical info and context): no effect on sensitivity or specificity
Elmore (1994)(108) **Study design** Real life: experimental **Objective** To investigate variability in radiologists' interpretations of mammograms **Type of analysis** statistical	Using a technique of stratified random sampling, 150 mammograms obtained in 1987 were selected: 27 from women with histopathologically confirmed breast cancer and 123 from women with no evidence of breast cancer after 3 years of follow-up examinations. Ten radiologists, who were unaware of the diagnoses and research hypothesis, each interpreted the 150 mammograms. Disagreement was analysed within pairs of the 10 radiologists as for the group of 150 women as a whole.	Observer variation	The diagnostic consistency between pairs of radiologists was moderate, with a median weighted percentage of agreement of 78% (weighted kappa 0.47). The frequency of radiologists' recommendations for an immediate workup ranged from 74 to 96% for mammograms from the women with cancer and from 11-65% for films from the women without cancer. A substantial disagreement in management recommendations occurred in 3% of the pairwise comparisons but in 25% of the comparisons for the group of women as a whole.
Elmore (1997)(109) **Study design** Real life: experimental **Objective** To determine whether mammographic interpretations are biased by the patient's clinical history **Type of analysis** statistical	On 2 occasions, separated by a 5 month wash-out period, 10 radiologists read mammograms for the same 100 women, randomly divided into 2 groups of 50. For 1 group, the clinical history was supplied for the first reading and omitted (except for age) for the second reading. This sequence was reversed in the other group. In addition, 5 cases were shown a third time with a deliberately leading sham history. 64 patients had mammographic abnormalities and 18 had breast cancer.	Clinical review bias	Knowledge of the clinical history altered the radiologists level of diagnostic suspicion and overall diagnostic accuracy did improve. Changes were made towards appropriate further diagnostic workup: an alerting history (e.g. breast symptoms or family history of breast cancer) increased the number of workups recommended in patients without cancer ($p=0.01$) and a nonalerting history led to fewer recommended workups in the cancer patients ($p=0.02$). The direction of the sham histories led an average of 4 of the 10 radiologists to change previous diagnoses and an average of 1 radiologists to change a previous biopsy recommendation.

Study details	**Methods**	**Bias**	**Evidence provided***
Erly(2003)(110) **Study design** Real life: prospective diagnostic accuracy study **Objective** To assess the accuracy of general radiologists in the interpretation via teleradiology of emergency CT scans of the heard **Type of analysis** Statistical	716 consecutive CT scans were interpreted by group of 15 general radiologists practicing in the community. Each CT scan was also examined by one of five neuroradiologists (gold standard) in an academic setting.	Observer Variation	Radiologist vs. neuroradiologist: decreased sensitivity no effect on specificity
Ewald(2006)(121) **Study design** Numeric: modelling **Objective** To examine the extent of bias introduced by the use of post hoc data driven analysis to generate an optimal diagnostic cut point for each data set. **Type of analysis** Statistical	Analysis of simulated data sets of test results for diseased and nondiseased subjects. Thresholds for the analysis were generated by searching for the threshold that gave the greatest sum of sensitivity and specificity and comparing this to the results from the prespecified threshold of 40.	*Threshold selection*	*Effect of data -driven threshold compared to pre-specified threshold: increased sensitivity and specificity. Size of bias decreases with increasing sample size but is also affected by the size of the smallest group so large samples with low disease prevalence can be affected.*
Froelicher (1998)(77) **Study design** Real life: diagnostic accuracy study, prospective **Objective** To compare the diagnostic utility of empirical scores, measurement and equations with that of visual ST-segment measurement in patients with reduced workup bias. **Type of analysis** statistical	Consecutive patients presenting with angina pectoris were recruited. Digital electrocardiographic recorders and angiographic callipers were used for testing. Sensitivity and specificity was calculated and compared to other similar studies conducted in populations where workup bias was present.	Test technology	No difference was found between computerised readings and physician readings.
		Partial verification bias	Standard exercise tests had lower sensitivity but higher specificity in this population with reduced work-up bias than in previous studies.
		Clinical review bias	The provision of additional information was found to improve test performance.
Gaffkin(2010)(35) **Study design** Real life: prospective diagnostic accuracy study **Objective** To show how the assumptions needed for unbiased statistical adjustment for verification bias can by undermined by conditions on the ground, and that accuracy of estimates is also compromised by too low a sampling fraction of subjects who test negative. **Type of analysis** Statistical	The accuracy of visual inspection with acetic acid (VIA) (index test) was compared to colposcopy (reference standard) for screening for cervical cancer. In Phase I all women testing positive and 10% random sample of those testing negative were assessed using the reference standard. However, study protocol was not followed and not all of those negative referred for biopsy received the test. In Phase II 2182 women were enrolled and all received both index test and reference standard on the same day.	Demographic Features	History of sexually transmitted diseases: no effect on sensitivity or specificity
		Prior testing	Pap test status: no effect on sensitivity or specificity
		Partial verification	1. Phase I (verification bias, not meeting missing at random assumption) vs. Phase II (no verification bias): decreased sensitivity and specificity 2: Adjustment for verification bias using the Begg and Green method lead to an overestimate in specificity and a considerable underestimate of sensitivity

Study details	**Methods**	**Bias**	**Evidence provided***
Geleijnse(2009)(36) **Study design** Real life: review **Objective** To assess the influence of various potential sources of bias on the diagnostic accuracy of dobutamine stress echocardiography (DSE). **Type of analysis** Statistical	62 studies of DSE were included (n=6881). Summary sensitivity and specificity were estimated for all studies combined and stratified according to various potential sources of bias	Demographic Features	History of MI: increased sensitivity no effect on specificity. Other patient related factors (medication use, age, gender) showed no association.
		Disease Severity	Extent of CAD (multivessel vs. single vessel involvement): increased sensitivity no effect on specificity
		Distorted Selection of participants	Pre-test CAD probability: increased sensitivity, decreased specificity; inclusion of patients with rest wall motion abnormalities: no effect on sensitivity or specificity
		Test execution	Quantitative scoring of CAG: no effect on sensitivity or specificity
		Test Technology	Older vs. newer technology: no effect on sensitivity or specificity
		Partial verification	Presence of referral (partial verification) bias: no effect on senility, decreased specificity
		Review Bias	Blind reading of reference standard or index test (was blinded in all but 5 studies): no effect on sensitivity or specificity
Gilbert(2002)(37) **Study design** Real life: review **Objective** To account for variation in test characteristics between studies of EEG accuracy **Type of analysis** Statistical	25 studies of accuracy of EEG to predict seizure recurrence were included. The influence of readers' thresholds for classifying EEG as positive, pre-test probability, proportion of patients with prior neurologic impairment, proportion treated and years followed were investigated using linear regression based on Moses-Littenberg with the percentage explained variance as the main outcome.	Demographic Features	Proportion of remote symptomatic patients, proportion of treated patients: no effect on overall accuracy
		Disease Prevalence	Sample probability of seizure recurrence: no effect on overall accuracy
		Inappropriate reference standard	Years followed (reference standard consisted of clinical follow-up): no effect on overall accuracy
		Observer Variation	Threshold for interpreting a positive EEG: associated with overall accuracy
Good (1990)(111) **Study design** Real life: experimental **Objective** To examine the effects that a concise, objective, and potentially computer-extractable history would have on diagnostic accuracy in the interpretation of chest radiographs. **Type of analysis** statistical	A computerised patient-history form that could be integrated realistically into the clinical environment was developed. A series of studies in which 247 posteroanterior normal (79) and abnormal (168) chest radiographs were interpreted by four board-certified radiologists, both with and without accompanying clinical histories were performed. The radiologists recorded their confidence rating of the presence or absence of one or more of the following abnormalities: interstitial disease, nodule, and pneumothorax.	Clinical review bias	Analysis of receiver operating characteristics showed that, with the exception of interpretation of one abnormality by one radiologist, there were no statistically significant difference (p<0.05) between cases interpreted with and without the history form for any of the radiologists. Knowledge of clinical history in a concise objective and potentially computer extractable way did not improve the accuracy of chest radiograph interpretations for the detection of interstitial disease nodules and pnemothoraces.
Gupta(2003)(89) **Study design** Real life: review	The results of three studies that reported on the test characteristics of PSA were compared. Approximate verification bias corrections (adjusting based on	Partial verification	Effect of partial verification bias: increased sensitivity, decreased specificity
		Differential verification	Effect of differential verification where unverified test negative results were included in the 2x2 table as true negative results: increased sensitivity and specificity

Study details	Methods	Bias	Evidence provided*
Objective To review how verification and incorporation biases influenced studies assessing the performance of PSA **Type of analysis** Statistical	previously reported detection rate of 22% in the PSA range of 2.5 o 4ng/ml) were applied to estimates of sensitivity and specificity stratified according to age and race. To adjust for incorporation bias, PSA was removed from the criteria establishing the absence of prostate cancer and the test characteristics of PSA was recalculated.	Incorporation	Effect of incorporation bias: increased sensitivity, decreased specificity
Haines(2007)(38) **Study design** Real life: review **Objective** To investigate design-related bias in hospital fall risk screening tool predictive accuracy evaluations **Type of analysis** Statistical	35 studies reporting 51 evaluations of risk screening tools were included in the review. The association between study design classification and the Youden index was assessed using linear regression with clustering based on screening tool.	Distorted Selection of participants	Trend for greater accuracy in prospective temporal design vs. prospective (external) designs (p=0.18). Authors used a non-standard definition of prospective. In addition to the typical definition, an a priori defined cut-off was required to be classified as prospective.
		Review Bias	Staff blinding: no effect on accuracy
		Sample size	No effect on accuracy
Hall(2004)(39) **Study design** Real life: retrospective diagnostic accuracy study **Objective** To assess whether spectrum bias is present in the evaluation of the diagnostic accuracy of rapid antigen detection test RADT compared to culture (reference standard) among children who are evaluated for pharyngitis. **Type of analysis** Statistical	Laboratory and clinical records from 561 consecutive children who underwent RADT were reviewed retrospectively. Patients were stratified according to the number of clinical features present using modified Centor criteria.	Disease Severity	Increasing Centor criteria increased sensitivity but no effect on specificity.
Hlatky (1984)(40) **Study design** Real life: diagnostic accuracy study, prospective **Objective** To investigate factors affecting the sensitivity and specificity of exercise electrocardiography **Type of analysis** statistical	Patients who had undergone both exercise electrocardiography and cardiac catheterisation. The effects on sensitivity of factors from clinical history, catheterisation, and exercise performance were defined by multivariable logistic regression analysis in 1401 patients with coronary disease; effects on specificity were defined by a similar analysis in 868 patients without coronary disease.	Demographic features	Five factors had significant independent effects on exercise electrocardiographic sensitivity: maximal exercise heart rate, number of diseased coronary arteries, type of angina and the patient's age and sex. Only maximal exercise heart rate had a significant, independent effect on exercise electrocardiographic specificity.

Study details	**Methods**	**Bias**	**Evidence provided***
Irwig(2006)(112) **Study design** Real life: experimental **Objective** To compare the combined accuracy of prior information and a test read with and without knowledge of prior information. **Type of analysis** Statistical	A study of cancer detection in women presenting with breast symptoms in whom ultrasound was read with and without reviewing prior mammography. A more sophisticated method for comparing the two sets of data using area under the curve is proposed and compared to results obtained using naïve analysis.	Clinical Review Bias	Interpretation of ultrasound with mammography on view was similar to interpretation without mammography based on the proposed methods; there was a difference based on naïve analysis (AUC higher when US interpreted with info on mammography).
Kittler(2002)(41) **Study design** Real life: review **Objective** To assess the influence of study characteristics on the accuracy of melanoma diagnosis with and without dermoscopy **Type of analysis** Statistical	27 studies were included in the review, DOR compared for assessment of melanoma without dermoscopy, with dermoscopy interpreted by experienced examiners and with dermoscopy interpreted by non-experts. Influence of study characteristics on the DOR investigated using univariate and multivariate SROC regression analysis.	Disease Prevalence	Increased prevalence: decreased DOR
		Review Bias	Test review bias: no association with DOR
		Observer Variation	Dermoscopy interpreted by expert greater DOR than when interpreted by non-expert examiners; dermoscopy more accuracy when interpreted by group of 2 or more experts vs. single interpretation
		Instrument Variation	Accuracy of dermoscopy for experimental studies that used presentation of slides, colour prints, or digital images lower DOR than for clinical studies in which diagnosis was made face to face
Lachs (1992)(42) **Study design** Real life: diagnostic accuracy study, prospective **Objective** To determine if the leukocyte esterase and bacterial nitrite rapid dipstick test for urinary tract infection (UTI) is susceptible to spectrum bias. **Type of analysis** statistical	PATIENTS: A total of 366 consecutive adult patients in whom clinicians performed urinalysis to diagnose or exclude UTI. SETTING: An urban emergency department and walk-in clinic. MEASUREMENTS: After the patient encounter, but before dipstick test or culture was done, clinicians recorded the signs and symptoms that were the basis for suspecting UTI and for performing a urinalysis and an estimate of the probability of UTI based on the clinical evaluation. For all patients who received urinalysis, dipstick tests and culture were done in the clinical microbiology laboratory by medical technologists blinded to clinical evaluation. Sensitivity for the dipstick was calculated using a positive result in either leukocyte esterase or bacterial nitrite, or both, as the criterion for a positive dipstick, and greater than 10(5) CFU/ml for a positive culture.	Disease prevalence	RESULTS: In the 107 patients with a high (greater than 50%) prior probability of UTI, who had many characteristic UTI symptoms, the sensitivity of the test was excellent (0. 92; 95% CI, 0.82 to 0.98). In the 259 patients with a low (less than or equal to 50%) prior probability of UTI, the sensitivity of the test was poor (0.56; CI, 0.03 to 0.79). Specificity in these two groups was 0.42 (0.28 – 0.57) and 0.78 (0.73-0.79) respectively. CONCLUSIONS: The leukocyte esterase and bacterial nitrite dipstick test for UTI is susceptible to spectrum bias, which may be responsible for differences in the test's sensitivity reported in previous studies. As a more general principle, diagnostic tests may have different sensitivities or specificities in different parts of the clinical spectrum of the disease they purport to identify or exclude, but studies evaluating such tests rarely report sensitivity and specificity in subgroups defined by clinical symptoms.

Study details	Methods	Bias	Evidence provided*
Lauer(2007)(90) **Study design** Numeric: modelling **Objective** To study the impact of verification bias on the estimated accuracy of PET in suspected lung cancer. **Type of analysis** Statistical	534 consecutive patients referred for PET were included. The accuracy of PET was evaluated against the gold standard of tissue acquisition and two methods (Begg and Greenes 1983 method as described by Miller; Diamond 1986 and 1993 method) were used to correct for verification bias.	*Partial verification*	*Impact of verification bias for cancer of any site increased sensitivity and decreased specificity; Impact of verification bias on PET for detection of mediastinal cancer: no association*
Leeflang(2008)(122) **Study design** Numeric: modelling **Objective** To determine the magnitude of bias in sensitivity and specificity associated with data driven selection of cut-off values and to examine potential solutions to reduce this bias. **Type of analysis** Statistical	Different sample sizes, distributions, and prevalences were used in a simulation study. Data-driven estimates of accuracy based on the Youden index were compared with the true values and the median bias was calculated. Three alternative approaches (assuming a specific distribution, leave-one-out, smoothed ROC curve) were examined for their ability to reduce this bias.	*Threshold selection*	*Data driven optimisation of threshold overestimates accuracy. Magnitude of bias greater with smaller sample sizes. More robust methods were less prone to bias.*
Leeflang(2009)(43) **Study design** Real life: review **Objective** To identify and explore mechanisms that may be responsible for sensitivity and specificity varying with prevalence. **Type of analysis** Statistical	Mechanisms that may be responsible for variations in estimates of sensitivity and specificity with prevalence are discussed and illustrated with examples from the literature	Disease Prevalence	Direction and magnitude of effect varied across studies
Levy (1990)(44) **Study design** Real life: diagnostic accuracy, prospective **Objective** To examine the sensitivity and specificity of the ECG as a tool for detecting electrocardiographically defined LVH (left ventricular hypertrophy) in a population based sample and to examine the impact of a variety of factors that attenuate the sensitivity and specificity of the ECG for the detection of LVH. **Type of analysis** statistical	Electrocardiographic criteria for LVH were examined in 4684 subjects of the Framingham Heart Study who underwent echocardiographic study for LVH. The chi-squared test was used to test for differences between sexes in the sensitivity and specificity of the ECG for echocardiographically defined LVH. The Cochran-Mantel-Haenszel statistic was used to adjust for sex and test the association between cigarette smoking and sensitivity and specificity of the ECG. Bivariate logistic regression was used to adjust for sex and test the sensitivity and specificity trends with increasing age, obesity and left ventricular mass/height.	Demographic features	Influence of sex: sensitivity was marginally lower in women (5.6 vs. 9%, p=0.075)specificity was high in both sexes (99.4% in women and 98.1% in men). Influence of age: There was a trend for sensitivity to increase with increasing age (p<0.0001, sex adjusted), there was a trend for specificity to decline with advancing age (p<0.001, sex adjusted) Influence of obesity: sensitivity was inversely related to increasing body mass index (p<0.05 for trend, sex adjusted), no specific differences in specificity was observed Influence of smoking: sensitivity was lower among smokers compared to non-smokers(5.7% v 10.9% in women, 1.6% v 8% in women, p=0.001 sex-adjusted). There were no statistically significant differences in specificity.

Study details	Methods	Bias	Evidence provided*
		Disease severity	Influence of severity of left ventricular hypertrophy: a statistically significant trend towards increasing sensitivity of the ECG with increasing severity of LVH was observed for both sexes (p<0.001).
Lijmer (1999)(18) **Study design** Real life: Meta-review **Objective** To empirically determine the quantitative effect of study design shortcomings on estimates of diagnostic accuracy **Type of analysis** statistical	Observational study of the methodological features of 184 original studies evaluating 218 diagnostic tests. Meta-analyses on diagnostic tests were identified through a systematic search of the literature. Association between study characteristics and estimates of diagnostic accuracy were evaluated with a regression model. The relative diagnostic odds ratio, which compared the diagnostic odds ratios of studies of a given test that lacked a particular methodological feature with those without the corresponding shortcomings in design was used as the outcome measure.	Demographic features	Diagnostic performance was overestimated when no description of the population under study was provided (RDOR, 1.4, 95% CI: 1.1, 1.7)
		Distorted selection of participants	Studies evaluating tests in a diseased population and a separate control group overestimated the diagnostic performance compared with studies that used a clinical population (RDOR, 3.0 95% CI: 2.0-4.5). Non-consecutive patient enrolment did not have any significant effect on diagnostic performance (RDOR: 0.9, 95% CI: 0.7, 1.1) neither did a retrospective study design (RDOR 1.0, 95% CI: 0.7, 1.4).
		Test execution	When no criteria for the test were described diagnostic performance was overestimated (RDOR 1.7, 95% CI: 1.1-2.5). When no criteria for the reference standard execution were described diagnostic performance was underestimated (RDOR: 0.7, 95% CI: 0.6, 0.9).
		Partial verification bias	Partial verification (when more than 10% of the study group did not receive the reference standard) was not associated with diagnostic performance (RDOR, 1.0, 95% CI: 0.8-1.3).
		Differential verification bias	Studies in which different reference standards were used for positive and negative results of the test under study overestimated the diagnostic performance compared with studies using a single reference standard for all patients (RDOR, 2.2 95% CI: 1.5-3.3).
		Review bias	Diagnostic performance was overestimated when the reference standard was interpreted with knowledge of the test result (RDOR, 1.3, 95% CI: 1.0-1.9).
Lijmer (1996)(91) **Study design** Real life: diagnostic accuracy, retrospective **Objective** To investigate the diagnostic accuracy of selected non-invasive tests for assessing peripheral arterial disease and to examine verification bias. **Type of analysis** statistical	Results of non-invasive tests in patients aged 40+ performed for suspected peripheral arterial disease were retrieved retrospectively from a computerised database. All angiograms (reference standard) performed within 2 months of the non-invasive tests were retrieved. Data were retrieved for 464 consecutive patients. The non-invasive test results warranted angiography in only 53 (12%) of the 441 patients studies, the other patients had milder forms of peripheral arterial disease and were therefore subjected to exercise training, counselling and follow-up. The estimates were corrected for verification bias using the method of Begg and Greenes 1983.[63]	Partial verification bias	The individual operating points on the ROC curves shifted after correcting for verification bias. For any particular threshold values, both true- and false- positive ratios changed after correcting for verification bias and the corrected likelihood ratio was closer to 1.0 than the likelihood ratio calculated from the verified sample.

Study details	**Methods**	**Bias**	**Evidence provided***
Mastandrea(2008)(45) **Study design** Real life: review **Objective** To investigate sources of heterogeneity affecting BNP for assessing heart failure severity. **Type of analysis** statistical	67 studies (98 samples) were included in the review. DORs were pooled using the DerSimonian and Laird random effects model. ANOVA was used to investigate the association of various possible sources of heterogeneity with the DOR.	Demographic Features	Age, sex, BMI: no effect on DOR
		Disease Severity	Disease severity: associated with DOR
		Disease Prevalence	Disease prevalence: associated with DOR
		Reference standard	Reference Method: associated with DOR
		Instrument Variation	Laboratory method: no effect on DOR
		Threshold selection	Threshold selected to maximise accuracy vs. other method of threshold selection: no effect on DOR
Medeiros(2007)(46) **Study design** Real life: retrospective diagnostic accuracy and CC study **Objective** To assess the effects of study design and spectrum bias on the diagnostic accuracy of confocal scanning laser opthalmoscopy (CSLO) in glaucoma. **Type of analysis** Statistical	Analysis 1 included 67 eyes with glaucomatous visual field loss and 56 eyes of normal volunteers. Analysis 2 included a cohort of patients with suspected glaucoma (40 eyes with progressive glaucomatous optic disc change were included in the glaucoma group and 43 eyes without any evidence of progressive damage to the optic nerve were included in the normal group). Areas under the ROC curves (AUC) were used to evaluate accuracy and were compared between the two analyses.	Distorted Selection of participants	Case-Control versus retrospective cohort – increased AUC
Melbye (1993)(47) **Study design** Real life: diagnostic accuracy study, prospective **Objective** To study the influence of the spectrum of patients on the usefulness of five clinical cues <very annoying dyspnoea>, <strong lateral chest pain>, crackles, C-reactive protein analysis, and erythrocyte sedimentation rate, in the diagnosis of pneumonia **Type of analysis** statistical	The diagnostic properties (sensitivity, specificity, likelihood ratio and positive predictive value) of the cues compared to radiographic pneumonia were evaluated for the following groups: 1. All 581 included patients 2. In 402 patients who also underwent physical chest examination 3. In 188 patients classified by the doctors as having a lower respiratory tract infection 4. In 79 patients referred for radiography by the doctors Only 229 of patients had radiographs (reference standard) ordered by doctor or nurse, an additional 25% of the remaining patients were also referred for radiography, none of these had pneumonia and so it was assumed that none of the remaining patients had pneumonia.	Demographic features	The specificity of very annoying dyspnoea decreased with increasing prevalence of pneumonia from 0.94 to 0.79, the LR dropped from 5.7 to 2.0, for strong lateral chest pain the drop in specificity was smaller from 0.93 to 0.90. Crackles was the only finding with a marked increase in sensitivity from 0.35 to 0.58, specificity dropped from 0.91 to 0.60 and the LR from 3.7 to 1.4, the PPV was nearly unchanged as the prevalence of radiographic pneumonia increased. A marked drop in specificity from 0.97 to 0.89 and LR from 9.2 to 2.3 was demonstrated for ESR. There was little change in PPV. A different pattern of changes was found for CRP, specificity was lower in the total group than in the 402 auscultated patients and the 188 patients classified as having LRTI. A corresponding rise in LR from 3.7 to 6.7 was found, PPV increased from 0.12 to 0.43 through the four levels of selection.
Michaud(2002)(48) **Study design** Real life: review **Objective**	26 studies were included in the review. The association broncheolar lavage volume, patient selection and prior treatment with anitbiotics with accuracy was estimated by calculating Q* using SROC regression separately for	Demographic Features	Prior treatment: estimates of Q* varied according to antiobiotic exposure as did the relative accuracy of the different tests.
		Distorted Selection of participants	Appropriate patient selection: increased sensitivity and specificity

Study details	Methods	Bias	Evidence provided*
To investigate how study design and previous antibiotic exposure influence the accuracy of various diagnostic tests for diagnosis of ventilator-associated pneumonia. **Type of analysis** Statistical	each of four and stratified according to these three quality criteria.	Test Technology	Higher BAL volume: increased sensitivity and specificity
		Inappropriate reference standard	Use of diagnostic consensus criteria as reference standard: no effect on sensitivity or specificity
Miller (1998)(92) **Study design** Real life: diagnostic accuracy study, retrospective **Objective** To investigate the effect of adjusting for post-test referral bias **Type of analysis** statistical	15 945 patients without prior myocardial infarction or revascularisation who underwent stress T1-201 or Tc-99m sestamibi imaging, 1771 underwent coronary angiography within 3 months after perfusion imaging. Sensitivity and specificity were calculated for the angiographic subgroup and the entire study population using a statistical method (Diamond method) that adjusts for referral bias.	Partial verification bias	Post-test referral bias (workup bias) leads to an overestimation of sensitivity (estimated as 97%, 66% after mathematical correction) and an underestimation of specificity (estimates as 13%, corrected estimate 73%).
Miller(2002)(49) **Study design** Real life: retrospective diagnostic accuracy study **Objective** To evaluate the effect of referral bias on the accuracy of SPECT for the diagnosis of coronary artery disease. **Type of analysis** statistical	Retrospective analysis based on data from Mayo clinic database. 14 273 patients without known coronary artery disease underwent stress SPECT. Coronary angiography was performed within 3 months after the stress test in 1853 patients (13%). The apparent sensitivity, specificity, and likelihood ratios of SPECT were determined in these patients, and then adjusted for verification bias using two different formulas - Begg and Greenes (1983) which adjusts for both pre-test and post-test referral bias and Diamond (1986) which only adjusts for post-test referral bias.	Demographic Features	Gender: no effect on sensitivity or specificity
		Test Technology	Type of radio-isotope technique: no effect on sensitivity or specificity
		Partial verification	Impact of adjusting for verification bias using either method (results similar for both methods): decreased sensitivity, increased specificity
Mol (1999)(93) **Study design** Real life: review **Objective** To evaluate the effect of verification bias on the accuracy of first-trimester nuchal translucency measurement for Down syndrome detection **Type of analysis** statistical	MEDLINE and EMBASE were searched to identify all papers relating the results of nuchal translucency measurement to foetal karyotype. The detected studies were scored for verification bias. Fifteen studies without and ten with verification bias were included.	Partial verification bias	Sensitivity and specificity were calculated for each study. For studies with verification bias, adjusted estimates of the sensitivity were calculated assuming a foetal loss rate for Down syndrome pregnancies of 48%. The sample size weighted sensitivity was 55% in studies without and 77% in those with verification bias, for specificities of 96% and 97%, respectively. After adjustment for verification bias, the sample size weighted sensitivity changed from 77% to 63%. Studies with verification bias reported higher sensitivities, but also slightly higher specificities of nuchal translucency measurement than studies without verification bias.

Study details	Methods	Bias	Evidence provided*
Moons (1997)(50) **Study design** Real life: diagnostic accuracy study, prospective **Objective** To evaluate the relevance of the sensitivity, specificity and LR of a test in clinical diagnosis, particularly for the same population as that from which the measures are derived. **Type of analysis** statistical	295 participants consecutively referred by GPs for evaluation of chest pain. Patient history, physical examination, results from symptom limited exercise testing and coronary angiography to determine the presence of coronary artery disease and the number of diseased vessels were recorded in that order. Coronary angiography took place within 3 months of the exercise test. Two experienced cardiologists who were blinded to the patient's history and test results independently interpreted the angiograms. The sensitivity and specificity of the exercise test was compared across patient subgroups (patient history, physical examination, exercise test and underlying disease severity).	Demographic features	The sensitivity of the ST/HR depression substantially differed according to sex, expected workload, absolute achieved workload, and relative workload SBP at peak exercise. Variation over smoking, cholesterol level, and baseline SBP was less marked. The specificity differed according to sex, diabetes, baseline SBP and relative workload. Although sensitivity and specificity were conversely affected by most variables, the LR of the exercise test still varied over categories of sex, smoking, cholesterol level, baseline SBP, relative workload and SBP at peak exercise.
		Disease severity	The sensitivity of the ST/HR depression varied according to number of disease vessels. Variation across patients with non-specific and atypical angina compared with typical angina was less marked
Moore(2005)(113) **Study design** Real life: retrospective diagnostic accuracy study **Objective** To compare the accuracy of physical therapists, orthopaedic surgeons and nonorthopaedic providers on patients with musculoskeletal injuries referred for MRI **Type of analysis** Statistical	Retrospective analysis of 560 patients referred for MRI. Electronic review of each patient's radiological profile performed to determine agreement between clinical diagnosis and MRI findings	Observer Variation	Physical therapists and orthopaedic surgeons had increased accuracy compared to nonorthopaedic providers
Morise (1994)(52) **Study design** Real life: diagnostic accuracy study, prospective **Objective** To investigate whether sex discrimination explains the differences in test accuracy among men and women referred for exercise electrocardiography **Type of analysis** statistical	4467 patients with suspected coronary disease who underwent exercise electrocardiography were studied employing a method to assess sensitivity and specificity without angiography. 18% of patients also underwent angiography. As a substitute for angiography the method used a disease probability model estimate made with a previously derived algorithm using age, sex, symptoms, diabetes, cholesterol and peak exercise heart rate. Positive exercise ST criteria were >= 1mm horizontal/downsloping depression.	Demographic features	The unbiased estimates of sensitivity and specificity were higher in men than in women (sensitivity = 40% vs. 33%, specificity = 96% vs. 89%).
		Partial verification bias	Sensitivity was higher and specificity lower in both men and women who underwent angiography compared to the whole group of patients. The absolute differences in the sensitivity and specificity before and after debiasing were similar in men and women indicating that the magnitude of workup bias in men and women was equivalent

Study details	Methods	Bias	Evidence provided*
Morise (1995)(51) **Study design** Real life: diagnostic accuracy study, retrospective **Objective** To compare the sensitivity and specificity of exercise electrocardiography in biased and unbiased populations of men and women **Type of analysis** statistical	To assess for sex-related differences in post-test referral bias, we compared the accuracy of exercise electrocardiography in biased (coronary angiography only) and unbiased (all unselected) populations with possible coronary disease. A retrospective analysis of clinical and exercise test data from 4467 patients (788 who underwent angiography) was performed (2824 men and 1643 women). The accuracy of a positive exercise test result was assessed in the entire unbiased group with a method that used disease probability (derived with a logistic algorithm) rather than angiography results.	Demographic features	Sensitivity and specificity were significantly greater in men than in women with use of the biased or unbiased groups. The amounts that sensitivity decreased and specificity increased, was not different for men and women. Therefore, the accuracy of exercise electrocardiography is lower in women than men irrespective of whether a biased or an unbiased group is used. However, these differences cannot be explained on the basis of sex-related differences in post-test referral bias.
		Partial verification bias	When the results for the unbiased and biased groups were compared, the sensitivities for the unbiased group were significantly lower and the specificities were significantly higher than those of the biased group. These differences reflect the effects of post-test referral bias.
O'Connor (1996)(53) **Study design** Real life: diagnostic accuracy, prospective **Objective** To investigate whether within the population of suspected multiple sclerosis (MS) patients, there would be differences in MRI and evoked potential (EP) sensitivity and specificity between those with mild MS versus those with more severe clinical disease. **Type of analysis** statistical	303 patients with suspected MS were evaluated by a board-certified neurologist, then scanned with MRI. Two hundred four patients also received EP testing. The group was divided into "possible" and "probable" MS subgroups and sensitivity and specificity for MRI and EP were calculated separately for these subgroups and the differences between them investigated.	Disease prevalence	The sensitivity of MRI in patients with suspected MS was 58 percent with a false-positive rate of 9%. The overall sensitivity was 64% in the probable and 45% in the possible group. In the low pre-test probability group sensitivity was 20%, and it was 70% in the high pre-test probability group. These differences in sensitivity are statistically significant ($p < 0.03$). In contrast, the specificity between groups did not differ significantly. EP sensitivity was 69% in the high probability subgroup and 5% in the low probability subgroup. ($p < 0.01$).

Study details	Methods	Bias	Evidence provided*
Panzer (1987)(94) **Study design** Real life: diagnostic accuracy study, prospective **Objective** To explore the potential impact of workup bias in prediction research by comparing the abilities of early clinical findings to predict intracerebral haemorrhage in biased and unbiased samples of patients with stroke **Type of analysis** statistical	A database containing clinical information concerning 374 patients with stroke and focal deficits meeting specific inclusion criteria was developed. Patients had undergone a physical and neurologic examination and basic laboratory test which was used to classify the patients as having had a haemorrhage or infarction. The "reference standard" for diagnosis was a CT scan which all patients included in the database had received on a routine basis. To model workup bias a simulated population in which CT scanning was not performed routinely, but instead was performed only in the presence of 3 specific clinical predictors of haemorrhage (headache, vomiting and decreased mental status) was assembled. 170 patients who had at least one of the three findings comprised the biased sample, the remaining 195 patients were excluded from the study population. Sensitivity, specificity, and likelihood ratios were calculated for various clinical predictors in both the biased and unbiased samples.	Partial verification bias	The frequency of each of the three clinical predictors used to select the biased sample was increased in that sample, this lead to increased sensitivity and decreased specificity in the biased as compared with the unbiased sample. The frequency of findings clinically related to the selection variables was also higher in the biased sample, the frequency of findings commonly associated with haemorrhage stroke but not directly related to those used to select the biased sample, was not consistently affected. In the biased sample likelihood ratios for the findings used to select the sample were consistently smaller than the likelihood ratios in the unbiased sample, likelihood ratios for related findings were also decreased, results were inconsistent for unrelated findings.
Phelps (1995)(95) **Study design** Numeric: Modelling **Objective** To use Monte Carlo methods to analyse the consequence of having a criterion standard that contains some error when analysing the accuracy of a diagnostic test using ROC curves. **Type of analysis** statistical	The authors use Monte Carlo studies to define inaccurate diagnostic tests and inaccurate fuzzy reference standards by adding various amounts of random noise to the true reference standard results. They then estimated ROC curves using this synthetic "diagnostic test" data and as the reference standard either the truth or the fuzzy reference standard results that measures the truth with error. They then compared the estimated ROC areas to determine the consequences of having an imperfect reference standard and the possible gains from using methods to offset the inherent FGS bias.	Inappropriate reference standard	*The results show that:* *1. When diagnostic test errors are statistically independent from inaccurate reference standard errors, estimated test accuracy declines.* *2. When the test and the fuzzy reference standard have statistically dependent errors, test accuracy can become overstated.*
Philbrick (1982)(54) **Study design** Real life: diagnostic accuracy study, prospective **Objective**	The exercise tests performed on a consecutive series of 208 patients in a tertiary-care university hospital and a university-affiliated community hospital were prospectively surveyed. When a patient was scheduled	Distorted selection of participants	If patients were excluded for the following reasons commonly used by researchers (the presence of clinical conditions that may produce false-positives or false-negatives) 48% of the 208 patients enrolled in the study would have been excluded. This would overestimate the test performance.

Study details	Methods	Bias	Evidence provided*
To investigate reasons for the wide variation in formal studies of sensitivity and specificity indexes for the diagnostic efficacy of the graded exercise test for angiographically defined coronary disease. **Type of analysis** narrative	for an exercise test the ordering physician was contacted to complete an outline of the patient's clinical status, reasons for ordering the test, and any plans for coronary arteriography (the reference standard test). After the test results were available the physicians were again contacted to determine whether the exercise test results influenced the decision to perform angiography. No patients were excluded from the study. The authors then discuss reasons why some of the patients included in their study would not be included in a diagnostic evaluation study and the theoretical effect that this would have on the estimates of test performance.	Partial verification bias	The reduced group of 127 patients would be further reduced by the requirement that patients have an invasive angiographic test to provide a definitive diagnosis. Patients are not always chosen randomly to receive the definitive test. Of the 171 physicians who answered the questionnaire 20 were urged to have angiography, in 19 cases physicians reported that the stress test results influenced their decision: 112 of these tests were positive, one was negative and 7 were non diagnostic. In 7 other cases a negative stress test result influenced the physician not to recommend angiography. The results show that work-up bias would have preferentially enriched the study group with patients who had positive exercise test results and reduced the number of patients with negative test results. These effects of work-up bias spuriously increase the sensitivity and lower the specificity obtained from exercise test research. Of the 20 patients recommended for angiography, 14 would have been excluded from the study group because of ineligibility, consequently only 6 patients (3%) would have become part of a definitely diagnosed study group. These 6 patients would be the "tip of the iceberg" constituting the admitted population for a customary study investigating the diagnostic efficacy of exercise testing.
		Handling of indeterminate results	If technically unsatisfactory exercise test results were excluded the 31% of the 205 test results would be excluded. If all patients with either a clinical reason for exclusion or a test result regarded as ineligible for the study group and were removed from further consideration 62% would be excluded.
Philbrick(2003)(96) **Study design** Real life: review **Objective** To verify the presence and magnitude of bias associated with the gold standard for the d-dimer test. **Type of analysis** Statistical	6 studies that compared D-dimer to imaging of both thigh and calf and that also stratified results by thigh and calf location were included.	Inappropriate reference standard	Estimates based on thigh imaging alone (optimal reference standard) compared to combined imaging of thigh and calf (imperfect reference standard): increased sensitivity, decreased specificity
Potchen (1979)(114) **Study design** Real life: experimental **Objective** To investigate the effect of irrelevant or directive chief complaint cues on normal and abnormal films of high and low degrees of difficulty. **Type of analysis** statistical	36 practising radiologists were divided into three equal size groups. Group 1 received cues directed to the correct diagnosis on 28 of 56 test P-A chest films and irrelevant complaints on the remaining 28. Group II received cues reversed for the same films. Group III received no patient data. The films had been divided into high and low difficulty categories based on consensus data from previous readers.	Clinical review bias	The patients' chief complaint assisted markedly in the interpretation of difficult abnormalities. 67% of these were detected with direct cues while only 48% and 44% were detected with irrelevant and no cues respectively (p<0.05).

Study details	**Methods**	**Bias**	**Evidence provided***
Pretorius(2007)(55) **Study design** Real life: retrospective diagnostic accuracy study **Objective** To investigate whether use of colposcopy as the reference standard inflates the sensitivity of acetic acid-aided visual inspection (VIA) compared to endocervical curettage (ECC)**Type of analysis** Statistical	375 women who had positive self or physician collected tests for high-risk HPV or abnormal cervical cytology and had VIA followed by colposcopy with directed biopsies and endocervical curettage (ECC) were reviewed (8497 originally screened) . Women had been assessed using a variety of index tests (visual inspection, cytology using three different thresholds and HPV based on physician and self-tests) and were compared to the reference standards of ECC ("optimum") and colposcopic directed biopsy.	Disease Severity	Sensitivity for detection of CIN2 or worse when 0-2 quadrants (less severe disease) involved less sensitivity than 3-4 quadrants (more severe) for cytology testing. No difference in sensitivity for physician or self-test.
		Inappropriate reference standard	Sensitivity of VIA compared to sub-optimum gold standard high than when compared to optimum gold standard. Estimates of sensitivity were also higher for all other screening tests but results were not statistically significantly different between reference tests.
Punglia(2003)(56) **Study design** Real life DA and modelling **Objective** To assess the screening characteristics of the PSA measurement after correction for verification bias **Type of analysis** Statistical	6691 underwent PSA screening for prostate cancer, 705 (11%) underwent biopsy (reference standard). A mathematical model (Begg and Greenes, 1983) was used to correct for the effects of verification bias and AUC of ROC curves were compared for adjusted and unadjusted estimates.	Demographic Features	Age (> vs. <60 years) decreased AUC. Previous test results (abnormal DRE examination) showed no effect on accuracy after correcting for verification bias.
		Partial verification	*Impact of adjusting for verification bias: decreased sensitivity, increased specificity and increased AUC.*
Raab (2000)(116) **Study design** Real life: experimental **Objective** To investigate the effect of the presence or absence of clinical history on the diagnostic accuracy of bronchial brush specimen interpretation was determined **Type of analysis** statistical	97 bronchial brush specimens were selected retrospectively from cytology files. Each of the specimens consisted of two slides, all cases had histologic and clinical follow-up, 49 cases had benign follow-up results 48 had malignant follow-up. The cases were divided into 3 groups and twice circulated among the study participants. On the first circulation no clinical history provided, on the second circulation, 2-3 months later clinical history was provided. Clinical history included was sex, age, clinical findings (if any), and clinical suspicion of disease. Each observer scored each case as definitely benign, probably benign, possible malignant, probably malignant and definitely malignant.	Clinical review bias	If clinical history was provided there was an increase in the number of malignant diagnoses. For every observer the likelihood ratio for the benign category was lower with clinical history than without clinical history - I.e. a benign diagnosis more likely indicated that a benign lesion was actually present if clinical history was provided than if clinical history was not provided. For the other diagnostic categories, depending on the observer, the presence of clinical history had a variable affect. For example, for the malignant category, if clinical history was provided the likelihood ratio increased for 2 observers and decreased for 3 observers. For each observer the positive predictive value of a malignant diagnosis was similar if history was or was not provided. For each observer, the negative predictive value was always higher if clinical history was provided. The means that when history is provided observers are more accurate with the benign diagnostic category and are able to shift malignant diagnoses out of this category. The diagnostic accuracy, as assessed using a ROC curve, of all pathologists increased if history was provided. For the pooled data across all pathologists there was a statistically significant different (p<0.05) between the accuracy of the diagnoses with history and without history.

Study details	Methods	Bias	Evidence provided*
Raab (1995)(115) **Study design** Real life: experimental **Objective** To use the bronchial brush specimen as an example, show the utility of using the LR and ROC curve in the evaluation of qualitative diagnoses. **Type of analysis** statistical	100 bronchial brush specimens were selected retrospectively from cytology files. Each of the specimens consisted of two slides, all cases had histologic and clinical follow-up, 50 cases had benign follow-up results 50 had malignant follow-up. The cases were divided into 3 groups and circulated among the study participants.	Observer variation	The LR for individual diagnostic categories varied among observers resulting in different clinically malignant probabilities. Observer experience did not appear to play a role in overall diagnostic accuracy, except in the diagnosis of small cell carcinoma.
Ransohoff (1978)(57) **Study design** Real life: review **Objective** To determine why many diagnostic test have proved to be valueless after optimistic introduction into medical practice by reviewing a series of investigations. **Type of analysis** narrative	Published studies of the carcinoembryonic antigen (CEA) test in the diagnosis of colonic cancer and the nitro-blue tetrazolium test (NBT) in the diagnosis of bacterial infection were examined. After an optimistic introduction into the medical community both these tests proved to be disappointing for their originally intended uses. English-language medical journals were searched for 1969-1973 for articles on CEA and from 1968-1973 for articles on NBT. Papers that had no original data, fewer than 10 patients or studies in which tests were used for prognosis, staging, or management rather than diagnosis were excluded. There were 17 reports for CEA and 16 for NBT.	Disease severity	CEA: The three studies reporting high sensitivity did not classify patients by any staging systems and so did not indicate whether patients with localised disease had been examined.. In 7/14 studies reporting lower sensitivity patients were classified by a staging system and patients with localised disease. The sensitivity of the test was shown to be much higher for extensive disease than in localised disease. The comparison group of the one study with high specificity contained patients with other cancers and colonic diseases but the extensiveness of these ailments was not reported. In the other 16 studies with low specificity, 6 indicated that an appropriate spectrum of comparative disease had been included. NBT: A wide clinical spectrum was no used in any of the four studies reporting high sensitivity but was reported in 5 of the remaining 12 studies which found lower sensitivity. The clinical and co-morbid components of spectrum of patients did not seem to be responsible for any major problems.
		Partial verification bias	CEA: Work-up bias did not appear to cause any major problems of missed diagnosis of colonic cancers. NBT: only one of the 16 studies reported precautions to avoid work-up bias, this study found a low sensitivity.
		Review bias	CEA: Biases of diagnostic interpretation and test interpretation were probably not important because both the test for CEA and pathology specimens are interpreted relatively objectively. NBT: the NBT test is interpreted subjectively and has a high degree of observer variability. Three studies contained precautions against biased test interpretation and only two tried to avoid biased diagnostic interpretation, only one of these studies found a high specificity for the test and none found a high sensitivity.

Study details	Methods	Bias	Evidence provided*
Ransohoff (1982)(97) **Study design** Real life: review of two studies **Objective** To provide an empirical illustration of diagnostic workup bias **Type of analysis** narrative	Two major reports examine the utility of serum ferritin in detecting iron overload in relative of patients with hereditary hemochromatosis (HH), reference standard is liver biopsy. Investigators from Brisbane found that ferritin was elevated in 15/15 relatives with marked iron overload as indicated by a histologic grade 3+ or 4+ hepatic iron. However, investigators from Boston reported substantially different results - elevated serum ferritin was found in none of 7 relatives who had 3+ or 4+ hepatic iron by histologic grading. This study aims to identify and assess possible reasons for these divergent results.	Partial verification bias	In the Boson study 62 relatives in two families were evaluated: 45 were examined and 34 had liver biopsies, biopsies were performed on normal relatives and on relatives with serum iron greater than 140ug/100ml. In the Brisbane study 199 relatives in 43 families were evaluated, only a few members of each family had biopsies and the reason for biopsy appears to have been an abnormal serum test. It appears that in Brisbane only relatives with abnormal tests were biopsied and so relatives with increased liver iron stores but normal serum tests would not have been identified.
Roger (1997)(58) **Study design** Real life: diagnostic accuracy study, prospective **Objective** To determine the effects of sex and of test verification bias on the diagnostic performance of exercise echocardiography. **Type of analysis** statistical	3679 consecutive patients (1714 women, 1965, men) who underwent an exercise echocardiographic were studied. The observed sensitivity, specificity and correct classification rate were calculated among 340 patients (244 men, 96 women) who underwent angiography. To study the effect of test verification bias, sensitivity and specificity were estimated for all patients who underwent exercise echocardiography including those not referred to angiography.	Demographic features	After correction for verification bias, sensitivity was lower in women than men.
		Partial verification bias	The observed sensitivity exercise echocardiography was 78% in men and 79% in women, the observed specificity was 37% in men and 34% in women. After adjustment for test verification bias, sensitivity was 42% in men and 32% in women, specificity was 83% in men and 86% in women.
Ronco (1996)(117) **Study design** Real life: experimental **Objective** To estimate the sensitivity of cytologists in recognising abnormal smears. **Type of analysis** Statistical	61 women with histologically confirmed cervical intraepithelial neoplasia (CIN) identified through colpohistrological and cytolgic screening. New smears were taken from study participants just before treatment, mixed with routine preparations, interpreted by unaware cytologists and then blindly reviewed by a group of three expert supervisors who reached a consensus diagnosis.	Observer variation	Sensitivity of the cytologists was less than that of the supervisors - they correctly diagnosed 30/34 smears judged as positive by supervisors.
Rozanski (1983)(59) **Study design** Real life: diagnostic accuracy study, retrospective **Objective** To verify the dramatic temporal decline in specificity of exercise radionuclide	Although exercise radionuclide ventriculography was initially reported to be a highly specific test for coronary-artery disease, later studies reported a high false-positive rate. To verify this turnabout, responses in 77 angiographically normal patients were analysed; 32 were studied from 1978 to 1979 (the early period), and 45 from 1980 to 1982 (the recent period).	Disease prevalence	Most patients studied in the early period had normal responses (94 per cent for ejection fraction and 84 per cent for wall motion). In contrast, normal responses were less frequent in patients studied in the recent period (49 per cent for ejection fraction and 36 per cent for wall motion, P less than 0.001). The probability of coronary disease before testing was higher in these patients (38 vs. 7 per cent, P less than 0.001). The temporal decline in specificity is partly a result of a change in the population being tested (pre-test referral bias).

Study details	Methods	Bias	Evidence provided*
ventriculography and to determine its cause. **Type of analysis** Statistical		Partial verification bias	More patients studied in the recent period underwent radionuclide ventriculography before angiography (78 vs. 22 per cent, P less than 0.001), and more of these prior studies had abnormal results than those performed after angiography (55 vs. 6 per cent, P less than 0.0001). The temporal decline in specificity is partly a result of a preferential selection of patients with a positive test response for coronary angiography (post-test referral bias).
Rutjes(2006)(60) **Study design** Real life: Meta-review **Objective** To determine and compare the direction and magnitude of the effects of a number of potential sources of bias and variation on estimates of diagnostic accuracy. **Type of analysis** Statistical	31 meta-analyses (487 primary studies) of the diagnostic accuracy of tests with at least 10 primary studies without preselection based on design features were included. A multivariable metaepidemiologic regression model was used to investigate the direction and strength of the association of 15 study features on estimates of diagnostic accuracy.	Distorted Selection of participants	Severe cases and healthy controls increased estimates of accuracy (RDOR 4.9, 0.6-37.3); other CC designs had no effect. Selection based on referral for index test results decreased accuracy (0.5, 0.3-0.9), no influence of selection bias on other test results. No association between use/avoidance of limited challenge group. Some suggestion that non-consecutive (RDOR 1.5, 1.0-2.1) and random sampling increased accuracy (RDOR 1.7, 0.9-3.2) estimates compared to consecutive samples. Retrospective data collection increased accuracy (RDOR 1.6, 1.1-2.2)
		Disease Progression	Effect of time interval (adequate, inadequate, not reported): no association with DOR
		Treatment Paradox	Effect of treatment (withheld, given, not reported): no association with DOR
		Inappropriate reference standard	Single vs. composite reference standard: no association with DOR
		Partial verification	Partial verification bias: no association with DOR
		Differential verification	Differential verification bias increased accuracy (RDOR 1.6, 0.9-2.9)
		Incorporation	Some suggestion of increased accuracy in presence of incorporation bias (RDOR 1.4, 0.7-2.8)
		Review Bias	Double blinded vs. single/nonblinded vs. not reported: no association with DOR
		Threshold selection	Some suggestion that post hoc definition of threshold increased accuracy (1.3, 0.8-1.9)
Rutjes(2003)(61) **Study design** Real life: Meta-review **Objective** To examine the influence of study design features on estimates of sensitivity, specificity, and diagnostic odds ratio in a series of meta-analyses. **Type of analysis** Statistical	Meta-epidemiologic approach, including 49 meta-analyses with 705 primary studies, covering a wide range of clinical conditions and test comparisons. A bivariate multivariable regression model was used to estimate the relative change in sensitivity and specificity between studies with specific design features and studies of the same test without these design features. The design features evaluated were type of design, timing of data-collection, patient selection, test result interpretation, and verification procedure.	Distorted Selection of participants	Design including severe cases and healthy controls versus other designs: no effect on sensitivity, specificity or DOR (rsens 1.60 (0.83 to 3.10); Retrospective versus prospective: none on sens, spec, or DOR Not consecutive versus consecutive: none on sens, spec, or DOR
		Partial verification	Partial verification versus complete verification: no effect on sensitivity, specificity or DOR
		Differential verification	Differential verification vs. full verification: no effect on sensitivity, increased specificity and DOR
		Review Bias	Single or not blinded versus double blinded: no effect on sensitivity, specificity or DOR
Santana-Boado (1998)(62) **Study design** Real life: diagnostic accuracy study, prospective	702 consecutive patients without previous myocardial infarction were studied with SPECT. 163 had coronary angiography (select minority) and 539 did not (silent	Demographic features	In verified patients sensitivity was lower in men than in women, but no gender difference in sensitivity was present after correction for verification bias.

Study details	Methods	Bias	Evidence provided*
Objective To compare the diagnostic accuracy of SPECT between both sexes and assess the influence of analysing only the patients with coronary angiography instead of all the patients submitted to study. **Type of analysis** **Statistical**	majority). All patients underwent exercise stress testing and simultaneous dipyrimadole was administered in 32% of patients who did not achieve maximum predicted heard rates. Diagnostic accuracy of the test was calculated for the select minority. Then sensitivity and specificity were recalculated according to the Diamond criteria.	Partial verification bias	The biased estimates of sensitivity were 95% in men and 85% in women (p=0.01). After mathematical correction for verification bias the 'debiased estimates were 88% and 87%, respectively (p=ns). The initial values for specificity were 89% in men and 91% in women (p=ns). After correction these were 96% and 91% (p=ns)
Schreiber (1963)(118) **Study design** Real-life: experimental **Objective** To investigate whether knowledge of clinical history has a favourable effect on the radiologist's perception of abnormal findings. **Type of analysis** Statistical	100 posteroanterior chest films were selected to be examined by 11 readers. Cards bearing the patient's age, sex, race and history number were prepared. Each film was read twice by each of the 11 readers. At the first reading half of the films were accompanied by the clinical history cards, at the second reading the half were accompanied by the clinical history cards. Each film was classified as positive or negative. Films were treated as truly positive if they were rated as positive more than 17 (out of a total of 22) times. Films were treated as truly negative for those which were read as negative more than 17 times. Films which could not be classified in this way were reclassified by discussion, 8 films could not be classified as positive or negative and these were discarded as indeterminate. Of the 92 films included in the study 24 were considered positive and 68 negative.	Clinical review bias	On average there were a greater proportion of true positives when the films were interpreted with clinical history than without (p=0.04). On average the number of false negatives was higher without history (4.2) than with history (2.7) (p=0.02) and the number of false positives was also higher without (7.1) than with, although this was not significant (p=0.18)
Shoaibi(2009)(63) **Study design** Real life: prospective diagnostic accuracy study **Objective** To assess the accuracy and correlates of the cardiac troponin (I (cTnI) assay in the diagnosis of non-ST-segment elevation MI and to determine how accuracy varies with gender. **Type of analysis** Statistical	924 patients with possible myocardial ischemia were included and the accuracy of cTnI (index test) was evaluated against a standard MI definition (reference standard)	Demographic Features	Gender: no effect on sensitivity or specificity
Sonad(2001)(78) **Study design**	27 studies comparing MRI to a pathologic standard in patients with clinically limited prostate cancer were	Test Technology	Studies that used fast SE imaging compared to conventional SE imaging; <1.5T vs. 1.5T and other coil versus endorectal coil increased overall accuracy

Study details	Methods	Bias	Evidence provided*
Real life: review **Objective** To determine the effect of high magnetic field strength, use of endorectal coil, use of fast spin-echo (SE) imaging and study size on the accuracy of MRI for staging prostate cancer. **Type of analysis** Statistical	included. Subgroup analyses examined magnetic field strength, use of an endorectal coil, use of fast SE imaging, publication date, and study size.	Sample size	Size <30: increased accuracy
Sohler(2008)(64) **Study design** Real life: prospective diagnostic accuracy study **Objective** To assess whether racial bias influence diagnoses assigned to patients at discharge from their first psychiatric hospitalisation **Type of analysis** Statistical	All patients admitted for their first psychiatric hospitalisation who self-identified as being black or white were included (n=491). Hospital clinical diagnosis (index test) was compared to interview based diagnosis based on DSM-III-R (reference standard).	Demographic Features	Estimates of accuracy in black vs. white patients: no effect on sensitivity or specificity
Stein (1993)(65) **Study design** Real life: diagnostic accuracy study, retrospective **Objective** To test the hypothesis that stratification of patients according to the presence or absence of prior cardiopulmonary disease may enhance the ventilation/perfusion scan assessment of PE among both clinical categories of patients **Type of analysis** statistical	Data were derived from an existing studies. Ventilation/perfusion lung scans were evaluated in 378 patients with acute PE and 672 patients in whom suspected PE was excluded. Patients were divided into two groups according to whether they had prior cardiac or pulmonary disease. Sensitivity, specificity and positive predictive value of PE based on the cumulative number of mismatched segmental defects were calculated separately for patients with and without cardiopulmonary disease. This data was stratified according to whether patients underwent obligatory angiography or patient requested angiography.	Disease severity	At >= 0.5 mismatched segmental equivalents positive predictive value was 80% among patients with no prior cardiopulmonary disease, compared to 68% in patients with prior cardiopulmonary disease (p<0.02), similar differences were seen for other numbers of mismatched segments. Sensitivity was higher in patients with prior cardiopulmonary disease than in those with prior cardiopulmonary disease at lower segmental equivalents but as segmental equivalents increased the difference decreased and sensitivity became higher in those with cardiopulmonary disease. Specificity was similar between the two groups. Areas under the ROC curve were higher for patients with no prior cardiopulmonary disease (0.8905 vs. 0.8215).

Study details	Methods	Bias	Evidence provided*
Steinbauer (1998)(66) **Study design** Real life: diagnostic accuracy study, prospective **Objective** To test for ethnic and sex bias in three self-report screening tests for alcohol use disorders in a primary care population **Type of analysis** statistical	Design: Study with primary care patients randomly selected from appointment lists. Setting: University-based family practice clinic. Patients: Probability sample of 1333 adult family practice patients stratified by sex and ethnicity. Measurements: Patients completed 1) a diagnostic interview to determine the presence of a current alcohol use disorder and 2) three screening tests: the CAGE questionnaire, the Self-Administered Alcoholism Screening Test (SAAST), and the Alcohol Use Disorders Identification Test (AUDIT)	Demographic features	The areas under the receiver-operating characteristic (ROC) curves for the CAGE questionnaire and the SAAST ranged from 0.61 to 0.88 and were particularly poor for African-American men and Mexican-American women. For the AUDIT, the area under the ROC curves was greater than 0.90 for each patient subgroup. The sensitivity of the CAGE questionnaire and the SAAST at standard cut-points was lowest for Mexican-American women (0.21 and 0.13, respectively). Positive likelihood ratios for the AUDIT were similar to or higher than those for the other screening tests, whereas negative likelihood ratios were lowest for the AUDIT (<0.33), indicating the superiority of this test in ruling out a disorder. A marked inconsistency in the accuracy of common self-report screening tests for alcohol use disorders was found when these tests were used in a single clinical site with male and female family practice patients of different ethnic backgrounds. The AUDIT does not seem to be affected by ethnic and sex bias.
Stengel(2005)(67) **Study design** Real life: review **Objective** To determine whether compliance with methodological standard affected the reported accuracy of screening ultrasonography (US) for suspected abdominal injury. **Type of analysis** statistical	An SR was conducted of prospective studies that compared US to any reference standard in patients with suspected abdominal injury. Studies were assessed using STARD and QUADAS. SROC and random effects meta-regression were used to model the effect of all methodological standards and other features on US sensitivity; specificity was consistently very high (pooled 99%) across studies. 62 studies were included.	Demographic Features	General population vs. children – increased sensitivity and specificity. No effect of including penetrating versus non penetrating injuries.
		Disease Severity	Mean injury severity score: no effect on sensitivity or specificity
		Distorted Selection of participants	Reporting of selection criteria; consecutive enrolment; prospective design: No effect on sensitivity
		Test execution	Reporting of methods of test execution (no effect on sens), fast vs. fast+ US (no effect for sens or spec)
		Test Technology	Higher transducer frequency: increased sensitivity
		Disease Progression	Reporting of time interval was associated with sensitivity; use of sufficiently short time interval showed no association
		Inappropriate reference standard	Use of single reference standard and reporting reference standard execution: decreased sensitivity
		Partial verification	Independent verification: decreased sensitivity
		Differential verification	Proportion of CT scans associated with sensitivity; proportion of laparotomies and proportion of diagnostic peritoneal lavage procedures no effect on sensitivity
		Review Bias	Blinding against US results, decreased sensitivity. Blinding against reference standard did not influence results.
		Observer Variation	Specification of sonography expertise and type of operatory (radiologist vs. surgeon): no effect on sensitivity
		Indeterminate Results	Handling of indeterminate results: no effect on sensitivity
		Withdrawals	Reporting of number of excluded patients and reporting of number of drop-outs: decreased sensitivity

Study details	**Methods**	**Bias**	**Evidence provided***
Syed(2008)(68) **Study design** Real life DA and modelling **Objective** To assess the accuracy of PET MPI for the diagnosis of angiographic CAD in patients without known CAD, corrected for verification bias. Secondary objectives were to evaluate accuracy in females and obese patients **Type of analysis** Narrative	833 PET studies performed in 122 patients without known CAD that were verified by coronary angiography within 3 months.. Results were adjusted for verification bias using 2 different models (Begg and Green and Diamond)	Demographic Features	Female vs. male: decreased sensitivity, increased specificity Obese vs. non-obese - effect varied: unadjusted sensitivity and specificity decreased, adjusted (Diamond), specificity decreased but no effect for sensitivity; adjusted Begg& Green no difference for sens or spec. No measure of statistical significance of association
		Partial verification	*Uncorrected (presence of verification bias) vs. corrected - direction of effect similar using both methods although actual estimates differed. Authors did not report any measure of statistical significance of differences.*
Taube (1990)(69) **Study design** Numeric: modelling with example using diagnostic accuracy design. **Objective** To demonstrate how possible selection mechanisms might influence the numerical sensitivity values **Type of analysis** statistical	Assume that a new method for detecting disease results in a measurement that increases with the development or severity of the disease. A simple model is presented which classifies the cases with the disease into three groups: 1. Those at an early stage of disease where the test will not be very effective e.g. mucinous 2. Those with fairly early disease in whom the test will be useful, the group relevant to the test. e.g. non-mucinous 3. Those with advanced disease in which it is obvious that they have the disease and for whom no screening device is necessary. e.g. Clearly malignant Sensitivities are then calculated for different combinations of these three groups using theoretical equations and also by using the example of a data-set of 168 cases of epithelial ovarian cancer.	Disease severity	The sensitivity calculated on all available data (i.e. for all three stages of disease combined) = 0.83 For the clearly malignant cases sensitivity = 0.96 For mucinous cases sensitivity = 0.46 For non-mucinous cases sensitivity = 0.87. However, if a proportion of non-mucinous cases cannot be sorted out by another method the future estimated sensitivity will be 0.74 *Theoretical simulations showed similar results to the example using data from epithelial ovarian cancer.*

Study details	Methods	Bias	Evidence provided*
Thibodeau (1981)(98) **Study design** Numeric: modelling **Objective** To evaluate the effect of misclassification by the reference standard on the observed sensitivity and specificity **Type of analysis** Statistical	Various statistical models were used to investigate how misclassification error may affect test performance.	Absent or inappropriate reference standard	In the case of conditional independence between the results of reference standard and diagnostic test, the observed sensitivity and specificity will be lower compared to the actual values if the reference standard contains error, as long as the diagnostic test is more often positive in the disease than in the non-diseased, and more negative in the non-diseased than in the disease. When conditional (positive) dependence is present between the reference and index test it would lead to lower values of observed sensitivity and specificity than would be obtained assuming independence.
Thompson(2006)(70) **Study design** Real life: retrospective diagnostic accuracy study **Objective** To determine the impact of finasteride on the accuracy of PSA for detecting prostate cancer. **Type of analysis** Statistical	The study included 4579 men receiving placebo and 5112 men receiving finasteride (participants in the prostate cancer prevention trial) who had a prostate biopsy and concurrent PSA tests during the 7-year study. For the placebo group the authors used commonly accepted PSA cut-offs; for the finasteride group, they used PSA cut-offs that were matched to obtain the same specicities as each cut-off in the placebo arm. Corresponding sensitivities and AUC of PSA were subsequently compared.	Demographic Features	Accuracy in men taking finasteride compared to men taking placebo. Increased AUC and sensitivity for all grades of cancer.
		Threshold selection	Fixed specificity in finasteride versus placebo arm: increased sensitivity
Tobin(2006)(71) **Study design** Real life: review **Objective** To determine the effects of spectrum and test-referral bias on the reported reliability of the frequency-to-tidal volume ratio (f/Vt) in predicting weaning success. **Type of analysis** statistical	Data updating an ACCP Task Force meta-analysis on sensitivity and specificity of f/Vt to predict weaning were extracted. Pre-test probability (prevalence) was used as an indirect measure of spectrum bias and verification bias. Authors evaluated if between study heterogeneity in sensitivity and specificity could be explained by pre-test probability (prevalence), as indicated by Chi-squared. Positive and negative predictive values were estimated for each study based on the pre-test probability of disease and sensitivity and specificity using Bayes theorem.	Disease prevalence	Increasing prevalence increases the positive predictive value and decreases the negative predictive value

Study details	Methods	Bias	Evidence provided*
van der Aa(2010)(119) **Study design** Real life: prospective diagnostic accuracy study **Objective** To determine the influence of knowledge of urine test outcome on the accuracy of cystoscopy (diagnostic review bias) during surveillance in patients with low grade, nonmuscle invasive urothelial carcinoma **Type of analysis** Statistical	Prospective RCT of surveillance by microsatellite analysis urine test in 448 patients. Urine test results were provided to the urologist in the intervention arm in which cystoscopy was done if the test was positive and at 3, 12 and 24 months. Urine tests results were not reported in the control arm in patients who underwent standard 3 month cystoscopy.	Review Bias	Diagnostic review bias: increased sensitivity
van der Schouw (1995)(72) **Study design** Real life: diagnostic accuracy, retrospective. **Objective** To investigate whether the differential diagnosis as registered directly in an existing data file could be used as an entrance to the indicated population. **Type of analysis** statistical	483 consecutive patients with clinical suspicion of scrotal pathology were enrolled in the study. Information on differential diagnoses, the final diagnosis and the ultrasonography results were available from the records of 372 patients who were included in the study. To investigate the values of the differential diagnosis as a potential entrance to the indicated population, patients were selected if they were suspected of having epididymitis according to their differential diagnosis, this resulted in a selection of 73 patients, by changing the criteria slightly a group of 108 patients was selected, by extending the criteria further a group of 183 patients were selected.	Disease prevalence	As the criteria used to select patients become stricter the test properties change markedly. As the selection criteria are widened (and so disease prevalence decreases), both sensitivity and specificity increase., the LR+ increased significantly from 4 to 28.
van Rijkom (1995)(73) **Study design** Real life: review **Objective** To investigate the influence of the diagnostic test, the study design and the validation method on reported validity. **Type of analysis** statistical	A systematic review was conducted. The sensitivity and specificity, study design (in vitro or in vivo experimental model) and the applied validation method were recorded. Validation methods were classified into two categories: strong and weak. D was calculated for each study. A multivariate analysis of variance with D as the dependent variable and diagnostic tests, validation methods and study design as independent variables was conducted. 39 sets of sensitivity and specificity were available.	Distorted selection of participants	On average values which originate from in vivo studies are higher than those from in vitro studies. In the multivariate analysis D values obtained from in vivo studies were significantly different from those obtained from in vitro studies (p<0.05), indicating that study design had a significant impact on the measurement of the validity of the diagnostic test.
		Inappropriate reference standard	On average weak validation methods yield higher values of D than strong validation methods. In the multivariate analysis D values were no t statistically significantly different between validation methods (p>0.05).
Wardlaw(2005)(120) **Study design**	15 studies on diagnosis of brain infarction from CT scans were included. Interobserver agreement was assessed	Clinical Review Bias	Knowledge of symptoms vs. no knowledge: no effect on sensitivity or specificity

Study details	Methods	Bias	Evidence provided*
Real life: review **Objective** To review CT signs in ischemic stroke to determine interobserver agreement and the relationship between early CT signs and patient outcome. **Type of analysis** Narrative	The analytical evaluation of the impact of some design and spectrum related items was omitted due to low number of studies included. Instead ranges of sensitivities and specificities were presented and discussed for subgroup of studies with and without optimal design choices.	Observer Variation	Experienced vs. less experienced observer: some suggestion that experienced observers performed better but insufficient data to formally investigate this.
Yoon(2009)(74) **Study design** Real life: prospective diagnostic accuracy study **Objective** To assess the effect of beta-blockers on global and per-vessel sensitivity and specificity of myocardial perfusion imaging (MPI) to identify significant and high-risk coronary artery disease using coronary angiography as the reference standard **Type of analysis** Statistical	555 patients underwent vasodilator MPI and had coronary angiography.	Demographic Features	Beta-blocker therapy versus no beta-blocker therapy: no effect on sensitivity or specificity
Zhang(2002)(75) **Study design** Real life: retrospective diagnostic accuracy study **Objective** To assess the influence of heart defect frequency and severity on screening sensitivity of the entire spectrum of congenital anomalies (CA) and on detection rate of congenital heart defect (CHD) when performing routine ultrasound screening in unselected pregnant women. **Type of analysis** Statistical	Secondary analysis was performed on prospective cohort data from Eurofetus, a large international collaborative study of ultrasound screening for CA in unselected populations and containing data for 3633 malformed foetuses. The following were assessed: frequency of CHD in the screened population and the global sensitivity of ultrasound in detecting CA (disease prevalence), association between the frequency of ventricular septum defect (VSD) and detection rate of CA and CHD (disease prevalence); and association between seriousness of CHD and CHD sensitivity (disease severity).	Disease Severity	Increased severity; associated CHD (CHD in presence of other malformations) versus single or multiple CHD: increase in sens; multiple versus single CHD: none on sens
		Disease Prevalence	Increased prevalence of CHD or VSD: decreased sensitivity

Study details	**Methods**	**Bias**	**Evidence provided***
Zhou (1994)(99) **Study design** Numeric: modelling with example using diagnostic accuracy design. **Objective** To examine the effect of verification bias on positive and negative predictive values **Type of analysis** statistical	The effect of verification bias on estimated positive and negative predictive values based on only patients with verified disease statuses (the so-called naïve estimators) were studied. By applying the maximum likelihood method the magnitude of the biases of the naïve estimators were quantified.	Partial verification bias	*Uses mathematical modelling to show that if the conditional independence assumption (that a patient's probability of selection for verification depends on only his/her test result) does not hold (i.e. if patient's probability of selection depends disease status) then the naïve estimators, estimated from only the verified patients, are biased.* Also presents an example of how this would work in practice. A total of 650 patients participated in a study. Of these 429 had a positive test result and 263 of these were referred to undergo disease verification procedures. Of the 221 patients with negative test results only 81 were referred to undergo disease verification procedures. The naïve estimators (using only verified cases) for the positive and negative predictive values are 88% (95% CI: 84-92) and 67% (95% CI: 57-77) respectively. The maximum likelihood estimators for the true range in positive and negative predictive values could range from 81-93% and 24-93% respectively. For this example the naïve estimator for the positive predictive values is reasonably robust against violation of the conditional independence assumption while the naïve estimator of the negative predictive value is sensitive to violation of the assumption.

* empirical evidence is reported in standard print, theoretical evidence is reported in italics.

Studies shaded in grey were included in the original bias and variation review(5;5)

Appendix 6: Summary of studies that have evaluated QUADAS

Study details and objective	Methods	Results	Recommendations for QUADAS
Bachmann(2009)(22) **Objective:** To study and formalise the fundamental mechanisms underlying spectrum and test review bias and to suggest amendments to STARD and QUADAS based on this.	Age, sex, cardiac symptoms and ECG results (index test) were recorded for 580 patient undergoing coronary angiography. The effects of different population compositions on the DOR of the ECG were investigated by simulating 100 hypothetical study populations with different proportions of patients with typical and atypical symptoms.	QUADAS recommends recording contextual information when interpreting a test but does not stipulate how to use this information when assessing test performance. QUADAS recommends evaluating the index test using the same clinical data available when using the test in practice. This does not exclude the possibility of variation in index test performance when using different sets of clinical data as there could be different views on what clinical data should be used in test evaluation. QUADAS insufficiently addresses the problems of selection and test review (clinical review) bias. Strict adherence to QUADAS does not preclude spectrum and test review (clinical review) bias.	QUADAS should be supplemented with an item addressing the appropriateness of statistical methods, in particular whether multivariable adjustments have been included in the analysis
Bauwens(2005)(123) **Objective:** To determine inter-rater agreement for QUADAS items and to clarify whether adherence to QUADAS affects measures of accuracy.	QUADAS was used to assess study quality in an SR of focused abdominal sonography for trauma, 62 studies were included. Two reviewers independently assessed studies using QUADAS and kappa statistics were used to measure agreement. Random effects meta-regression was used to model the impact of single QUADAS items on accuracy.	**Positive results:** All other items (10/14) showed substantial or almost perfect agreement. **Negative results:** Inter rater agreement was poor for appropriateness of the reference test (k=0.12), independence of the reference test (avoidance of incorporation bias) (k=0.03), specification or dropouts (0.23) and time interval between index and reference standard (k=0.24).	The appropriateness of the verification procedures must be defined according to the test under investigation (NOTE: unclear if this was specified a priori for this review).
Hollingworth(2006)(128) **Objective:** To assess the inter-rater reliability of QUADAS using data from an SR on MRS for the characterisation of suspected brain tumours.	19 DTA studies were included in the review and were independently assessed for study quality by two reviewers (from a pool of 6) using QUADAS. All reviewers were working at radiology departments and were specialized in neuroradiology and spectroscopy. 3 of them had previous experience in performing systematic review of diagnostic accuracy studies. Differences in reliability were compared with Fisher's exact test. The only change to the original QUADAS document was to replace the word "index test" with MRS. In addition, guidance was customised to the review where appropriate. If reviewers had questions regarding the QA this could be discussed with the rest of the group. Correlation, % agreement and kappa	**Positive results:** There was high agreement for the total number of items scored as yes for each study (rank correlation 0.78, p<0.01). The mean % agreement in rating individual QUADAS items was 69% and mean inter-rater reliability was 0.22 (unweighted kappa; fair agreement) ranging from -0.28 , no agreement beyond chance to 0.58, moderate agreement. Agreement was highest (84-90%) for items 1, 3, 5, moderate (60-80%) for items 2, 4, 6, 8, 9, 12 and 14 and lowest (47-58%) for items 7, 10, 11, and 13. There was no difference between reliability for validity items (% agreement 68%) versus generazability and reporting items (69%). QUADAS was found to be a very informative way of assess DTA study quality and the authors state that they would use it in future reviews.	The authors suggest the following to improve the reliability of QUADAS: 1. Ensure guidance is clear and adhered to for specific reviews 2. Provide individual feedback for reviewers in a pilot evaluation. 3. Extract in duplicate and resolve disagreements by consensus 4. It is not only description of the index test that is important but also the quality of the technique used which has an important impact on external validity, the following rewording is suggested: "Does the method used to perform the index test represent the current state of the art for that index test?". Similar wording for other items evaluating the clarity of reporting (items 2, 8, 9, 13 and 14) is suggested.

Study details and objective	Methods	Results	Recommendations for QUADAS
	statistics (weighted and unweighted) were recorded to measure inter-rater reliability. A prior hypothesis was that agreement between reviewers would be greater for the eight validity items than those for the generalizability (item 1) and clarity of study reporting (items 2,8,9,13,14).		5. The issue of time between index test and reference standard (item 4) is raised and the fact that it may not always be desirable/appropriate to have a short time interval between these. The authors agree with the recommendation that QUADAS should not be used to calculate a summary quality score. They also caution against the use of individual reviewers to assess study quality, especially if this is used as a basis for excluding studies or meta-regression analyses.
Lumbreras(2008)(124) **Objective:** To adapt QUADAS to the particular methodological challenges posed by new molecular diagnostic tests, and to fit QUADAS to each study phase, in order to contribute to the development of specific recommendation on genomics and proteomics (-omics)-based diagnostic research.	Five phases were used to adapt QUADAS to "omics" based diagnostic research - techniques that provide a comprehensive analysis of the (near)complete cellular specific constituents such as RNAs, DNAs, proteins and intermediary metabolites: 1. Preliminary decisions 2. Definition of phases 3. Preliminary item generation, including assessment of application of each QUADAS item to "omics" research 4. Evaluation of guidelines 5. Final generation of guidelines The new tool was named QUADRANOMICS	**Positive results:** An additional domain was added, where the extractor had to first indicate the design phase of the study, on a scale from 1 (healthy case-control study) to 4 (diagnostic cohort study). Thereafter, the assessment tool was applied. Most QUADAS items were retained exceptions were: Item 5: Independence of index test and reference standard as -omics based diagnostics tests are not currently used as the gold standard or part of the gold standard Item 14: Withdrawals, included in reformulated QUADRANOMICS item 1 Guidelines for scoring items were reworded for: Were selection criteria clearly described? Was the execution of the index test described insufficient detail to permit replication of the test? Items added were: Was the type of sample fully described? Were the procedures and timing of biological sample collection with respect to clinical factors described with enough detail? Sub questions: - Clinical and physiological factors - Diagnostic and treatment procedures Were handling and pre-analytical procedures reported in sufficient detail and similar for the whole sample? If differences in procedure were reported was their effect on the results assessed? Is it likely the presence of over fitting was avoided? Guidance on scoring items was provided. The scoring system developed for	None

Study details and objective	Methods	Results	Recommendations for QUADAS
		QUADAS of yes/no/unclear was retained and 'not applied' was added as an additional category.	
Mann(2009)(127) **Objective:** To examine the validity and usefulness of QUADAS when applied to DTA studies using psychometric instruments	Two reviewers independently assessed the quality of 54 studies assessing screening instruments for the diagnosis of post natal depression using QUADAS. QUADAS item 12 (availability of clinical information) was excluded as prior knowledge of clinical information was judged not to influence test results. For QUADAS item 4 a 2 week period was specified and item 13 was modified to refer to missing items/unclear responses rather than uninterpretable/intermediate results. The proportion of agreement between rates was calculated. In addition, the reporting of flow-diagrams was scored.	**Positive results:** The overall proportion of agreement between the two reviewers for all QUADAS items combined was 85.7%. The proportion of agreement between reviewers ranged from 57 to 100%. Agreement was good (>80%) for 8 items (3, 5, 6, 7, 8, 9, 10, 11). The poorest agreement was for patient spectrum, selection criteria, time between tests, uninterpretable test results, and withdrawals (items 1, 2, 4, 13, 14) - agreement ranged from 57% to 76% for these. Disagreement was generally between yes/unclear and no/unclear rather than between yes/no. Poor quality of reporting hampered quality assessment. Recommendations to improve reporting are provided in the paper. QUADAS was described as relatively easy to use **Negative results:** Items uninterpretable results and withdrawals caused problems in their application: the guidance notes were difficult to apply, especially in determining whether there were truly any withdrawals or uninterpretable results. In case-control designs, the authors found the scoring of partial verification difficult. To assess the full guidance with the modification based on the more recently suggest modifications two papers are required; it would be helpful to have complete guidance in a single location.	The clarify of the guidance for uninterpretable results and withdrawals should be addressed before application of the tool. Concluded that QUADAS was an acceptable tool for the quality appraisal of DTA studies using psychometric instruments to identify postnatal depression.
Meads(2009)(126) **Objective:** To describe modifications made to QUADAS to enable the assessment of diagnostic before-after studies, and to describe experience of using QUADAS, and its relation to published theory on diagnostic or therapeutic yield studies.	Two reviewers independently used QUADAS to assess study quality in a NICE rapid review of 24 studies on structural neuroimaging in psychosis. This review included diagnostic or therapeutic yield studies described as diagnostic before-after-studies: patients undergo existing test(s) and therapeutic strategy is noted, new test performed and any change in diagnosis/treatment strategy is noted. Design can be elaborated to include measurement of accuracy and assessment of patient outcome. For this review the	**Positive results:** The checklist allowed consistent and transparent assessment of quality of the included studies. **Negative results:** Items 3 and 7 were removed because only studies using CT or MRI were included in the review, and because in diagnostic before – after tests the 'after' diagnostic strategy (referred to here as the reference test) necessarily incorporated the 'before' component (referred to here as the index test), i.e. patients would not get CT/MRI alone. The following items were added: A. "Were patients recruited consecutively?" B. "Who performed the clinical evaluation and image analysis?" C. Was the study and/or collection of clinical variables conducted prospectively?	A better checklist would have provided more details about the spectrum of included patients. Suggested items include: duration of untreated disease; reason for referral of patients into the study; the setting of the study

Study details and objective	Methods	Results	Recommendations for QUADAS
	"before" diagnostic strategy was considered to be the index test and the "after" strategy was considered to be the reference standard. The outcomes of interest were diagnostic yield, therapeutic yield or clinical outcomes, and not diagnostic accuracy.	D. What was the explanation for patients who did not receive CT or MRI? - sub question of item 14 (withdrawals) The checklist did not lead to that much greater insight into the relationship between potential threats to validity identified by the checklist and the direction of results of the studies.	
Raatz(2010)(125) **Objective:** To determine whether QUADAS captured all relevant sources of bias when the index test was compared to a concurrent routine test and when the reference standard is follow-up	QUADAS applied in an SR of PET compared to conventional tests for assessing patients with lymphoma. The review included 7 studies, all used follow-up as the reference standard.	**Negative results:** QUADAS requests a short interval between index test and reference standard. With follow-up as the reference standard studies need to demonstrate sufficiently long follow-up to distinguish recurrence and healing. Reviewers need to assess the possibility of confounding during follow-up QUADAS evaluated the performance of the index test but not the comparator tests: It does not ask about mutual blinding of readers reviewing multiple tests It does not explore whether the statistical method takes into account the lack of independence of results of index and comparator tests when derived from the same patients.	A QUADAS update should consider additional criteria for situations in which a new index test is compared to a concurrent routine test and when the reference standard is follow-up
Whiting(2006)(2) Objective: To evaluate the validity and usefulness of QUADAS	Three reviewers independently rated the quality of 30 studies using QUADAS and 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. Reviewer 1 had previously carried out several diagnostic systematic reviews, had used QUADAS and 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. Variability was expressed as proportion of studies for which each reviewer agreed with the consensus rating. In addition, inter-observer variability was calculated by the kappa statistic. Twenty reviewers who had used QUADAS in	**Positive results:** **Over all items, the agreements between each reviewer and the final consensus** rating were 91%, 90% and 85%. Overall reviewer variability was good with a kappa of 0.65. The results for individual QUADAS items varied between 50% and 100% with a median value of 90%. 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. 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. In detail, 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. Reviewers typically omitted items on which no differences between studies were observed. 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	The evaluation highlighted particular difficulties in scoring the items on uninterpretable results and withdrawals. Revised guidelines for scoring these items were proposed. Major modifications to the content of QUADAS itself, in terms of items included, are not necessary. 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 a priori. Reviewers should consider whether all QUADAS items are relevant to review, and whether additional quality items should be assessed as part of their review. Clarity of phrasing should be a key consideration in a future revision.

Study details and objective	Methods	Results	Recommendations for QUADAS
	their reviews completed a short structured questionnaire on their experience of QUADAS.	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. There was substantial variation in the time taken to complete QUADAS, ranging from less than 10 minutes to over 1 hour **Negative results:** Items related to uninterpretable test results and withdrawals led to the most disagreements, followed by item: "Were selection criteria clearly described" and "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)". One surveyed reviewer felt that QUADAS 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. One reviewer indicated that the items availability of clinical information and withdrawals were difficult to score for case-control designs, and that in most cases the issue of follow-up was not relevant.	