Health Technology Assessment 2004; Vol. 8: No. 25

Development and validation of methods for assessing the quality of diagnostic accuracy studies

P Whiting, AWS Rutjes, J Dinnes, JB Reitsma, PMM Bossuyt and J Kleijnen

June 2004

Health Technology Assessment
NHS R&D HTA Programme

How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Development and validation of methods for assessing the quality of diagnostic accuracy studies

P Whiting,[1*] AWS Rutjes,[2] J Dinnes,[3] JB Reitsma,[2] PMM Bossuyt[2] and J Kleijnen[1]

[1] Centre for Reviews and Dissemination, University of York, UK
[2] Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, The Netherlands
[3] Wessex Institute for Health Research and Development, University of Southampton, UK

* Corresponding author

Declared competing interests of authors: none

Published June 2004

This report should be referenced as follows:

Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess* 2004;**8**(25).

Health Technology Assessment is indexed in *Index Medicus*/MEDLINE and *Excerpta Medica*/EMBASE.

NHS R&D HTA Programme

The research findings from the NHS R&D Health Technology Assessment (HTA) Programme directly influence key decision-making bodies such as the National Institute for Clinical Excellence (NICE) and the National Screening Committee (NSC) who rely on HTA outputs to help raise standards of care. HTA findings also help to improve the quality of the service in the NHS indirectly in that they form a key component of the 'National Knowledge Service' that is being developed to improve the evidence of clinical practice throughout the NHS.

The HTA Programme was set up in 1993. Its role is to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined to include all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care, rather than settings of care.

The HTA programme commissions research only on topics where it has identified key gaps in the evidence needed by the NHS. Suggestions for topics are actively sought from people working in the NHS, the public, consumer groups and professional bodies such as Royal Colleges and NHS Trusts.

Research suggestions are carefully considered by panels of independent experts (including consumers) whose advice results in a ranked list of recommended research priorities. The HTA Programme then commissions the research team best suited to undertake the work, in the manner most appropriate to find the relevant answers. Some projects may take only months, others need several years to answer the research questions adequately. They may involve synthesising existing evidence or designing a trial to produce new evidence where none currently exists.

Additionally, through its Technology Assessment Report (TAR) call-off contract, the HTA Programme is able to commission bespoke reports, principally for NICE, but also for other policy customers, such as a National Clinical Director. TARs bring together evidence on key aspects of the use of specific technologies and usually have to be completed within a limited time period.

Criteria for inclusion in the HTA monograph series
Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The research reported in this monograph was commissioned by the HTA Programme as project number 98/27/99. As funder, by devising a commissioning brief, the HTA Programme specified the research question and study design. The authors have been wholly responsible for all data collection, analysis and interpretation and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the referees for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

The views expressed in this publication are those of the authors and not necessarily those of the HTA Programme or the Department of Health.

ISSN 1366-5278

Published by Gray Publishing, Tunbridge Wells, Kent, on behalf of NCCHTA.
Printed on acid-free paper in the UK by St Edmundsbury Press Ltd, Bury St Edmunds, Suffolk.

M

Abstract

Development and validation of methods for assessing the quality of diagnostic accuracy studies

P Whiting,[1*] AWS Rutjes,[2] J Dinnes,[3] JB Reitsma,[2] PMM Bossuyt[2] and J Kleijnen[1]

[1] Centre for Reviews and Dissemination, University of York, UK
[2] Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, The Netherlands
[3] Wessex Institute for Health Research and Development, University of Southampton, UK
* Corresponding author

Objectives: To develop a quality assessment tool which will be used in systematic reviews to assess the quality of primary studies of diagnostic accuracy.
Data sources: Electronic databases including MEDLINE, EMBASE, BIOSIS and the methodological databases of both CRD and the Cochrane Collaboration.
Review methods: Three systematic reviews were conducted to provide an evidence base for the development of the quality assessment tool. A Delphi procedure was used to develop the quality assessment tool and the information provided by the reviews was incorporated into this. A panel of nine experts in the area of diagnostic accuracy studies took part in the Delphi procedure to agree on the items to be included in the tool. Panel members were also asked to provide feedback on various other items and whether they would like to see the development of additional topic and design specific items. The Delphi procedure produced the quality assessment tool, named the QUADAS tool, which consisted of 14 items. A background document was produced describing each item included in the tool and how each of the items should be scored.
Results: The reviews produced 28 possible items for inclusion in the quality assessment tool. It was found that the sources of bias supported by the most empirical evidence were variation by clinical and demographic subgroups, disease prevalence/severity, partial verification bias, clinical review bias and observer/instrument variation. There was also some evidence of bias for the effects of distorted selection of participants, absent or inappropriate reference standard, differential verification bias and review bias. The evidence for the effects of other sources of bias was insufficient to draw conclusions. The third review found that only one item, the avoidance of review bias, was included in more than 75% of tools. Spectrum composition, population recruitment, absent or inappropriate reference standard and verification bias were each included in 50–75% of tools. Other items were included in less than 50% of tools.
The second review found that the quality assessment tool should have the potential to be discussed narratively, reported in a tabular summary, used as recommendations for future research, used to conduct sensitivity or regression analyses and used as criteria for inclusion in the review or a primary analysis. This suggested that some distinction is needed between high- and low-quality studies. Component analysis was considered the best approach to incorporate quality into systematic reviews of diagnostic studies and this was taken into consideration when developing the tool.
Conclusions: This project produced an evidence-based quality assessment tool to be used in systematic reviews of diagnostic accuracy studies. Through the various stages of the project the current lack of such a tool and the need for a systematically developed validated tool were demonstrated. Further work to validate the tool continues beyond the scope of this project. The further development of the tool by the addition of design- and topic-specific criteria is proposed.

Contents

Glossary and list of abbreviations vii

Executive summary xi

1 Background 1
Evaluation methods for diagnostic tests 1
Systematic reviews of diagnostic tests 1
Quality of diagnostic studies 1
Quality assessment tools to assess diagnostic studies 2
Assessment tool validation 2
Using quality assessment in diagnostic test reviews 2
STAndards for Reporting Diagnostic Accuracy (STARD) project 3

2 Research questions 5

3 Approach 7

4 General methods 9
Classification of bias 9
Spectrum composition 9
Index test and reference standard 11
Interpretation 12
Analysis 13
Research planning (objective 3 only) 13

5 Objective 1: Review the literature on the concepts underlying diagnostic research and identify the main factors that can bias the results of diagnostic studies 15
Methods 15
Results 16
Discussion 26
Conclusion 29

6 Objective 2: Examine how quality assessment has been handled in systematic reviews ... 31
Methods 31
Results 32
Discussion 37
Conclusion 38

7 Objective 3: Examine existing methods or assessment tools that have been used to assess the quality of diagnostic research, and any evidence on which they are based 39
Methods 39
Results 40
Discussion 46
Conclusions 48

8 Objective 4: Develop a new evidence-based assessment tool for the quality assessment of diagnostic studies 49
Preliminary conceptual decisions 49
Item generation and assessment of item face validity 50
Delphi procedure 51

9 QUADAS background document 59
Background to the tool 59
Aims of the tool 59
The quality assessment tool 60
Explanation of items included in the quality assessment tool and guide to scoring items 60

10 Discussion and proposals for further work 67

Acknowledgements 69

References 71

Appendix 1 Search strategies 81

Appendix 2 Data extraction tables: objective 1 83

Appendix 3 Data extraction tables: objective 2 109

Appendix 4 Data extraction tables: objective 3 137

Appendix 5 Members of the advisory panel 185

Appendix 6 Delphi questionnaires 187

Health Technology Assessment reports published to date 235

Health Technology Assessment Programme 245

Glossary and list of abbreviations

Glossary: measures of diagnostic test performance

		Disease: Present	Disease: Absent
Test result	+	*a*	*b*
	–	*c*	*d*

True positives — Correct positive test result: *a* – number of diseased persons with a positive test result.[1]

True negatives — Correct negative test results: *d* – number of non-diseased persons with a negative test result.[1]

False positives — Incorrect positive test result: *b* – number of non-diseased persons with a positive test result.[1]

False negatives — Incorrect negative test result: *c* – number of diseased persons with a negative test result.[1]

Sensitivity — $a/(a + c)$; proportion of people with the target disorder who have a positive test result.

Specificity — $d/(b + d)$; proportion of people without the target disorder who have a negative test result.

Test accuracy — The proportion of test results correctly identified by the test: $(a + d)/(a + b + c + d)$

Likelihood ratio (LR): positive (LR +ve), negative (LR –ve) — Describes how many times a person with disease is more likely to receive a particular test result than a person without disease.[1] An LR of a positive test result is usually a number greater than 1; an LR of a negative test result usually lies between 0 and 1.

$$LR+ = \{a/(a + c)\}/\{b/(b + d)\}$$
$$= \text{sensitivity}/(1 - \text{specificity})$$
$$LR- = \{c/(a + c)\}/\{d/(b + d)\}$$
$$= (1 - \text{sensitivity})/\text{specificity}$$

Diagnostic odds ratio (DOR) — Used as an overall (single indicator) measure of the diagnostic accuracy of a diagnostic test. It is calculated as the odds of positivity among diseased persons, divided by the odds of positivity among non-diseased.[2] When a test provides no diagnostic evidence then the DOR is 1.0.[1]

$$DOR = \{a/c\}/\{b/d\}$$
$$= \{\text{sensitivity}/(1 - \text{specificity})\}/\{(1 - \text{sensitivity})/\text{specificity}\}$$
$$= LR\ +ve/LR\ -ve$$
$$= ad/bc$$

continued

Glossary: measures of diagnostic test performance *continued*

Predictive value

Positive predictive value: the probability of disease among all persons with a positive test result

Positive predictive value (PPV) $= a/(a + b)$

Negative predictive value: the probability of non-disease among all persons with a negative test result

Negative predictive value (NPV) $= d(c + d)$

Predictive values depend on disease prevalence: the more common a disease, the more likely it is that a positive test result is right and a negative result is wrong.[1]

Receiver operating characteristic (ROC)

A ROC curve represents the relationship between 'true-positive fraction' (sensitivity) and 'false-positive fraction' (1 – specificity). It displays the trade-offs between sensitivity and specificity as a result of varying the cut-off value for positivity in case of a continuous test result.[3]

Summary ROC curve

The summary ROC approach models test accuracy, defined by the log of the DOR (D = logit(sensitivity) – logit(1 – specificity)), as a function of test threshold (S = logit(sensitivity) + logit(1 – specificity)). S relates to the positivity threshold: it has a value of 0 in studies where sensitivity equals specificity; it is positive in studies where sensitivity is higher than specificity, and negative when specificity is higher than sensitivity. For a set of primary studies, the following linear regression model is fitted:

$$D = \alpha + \beta S$$

where D is the log odds ratio in each study, α is the intercept, which is the expected log odds ratio when $S = 0$, and β is the coefficient of S, indicating whether the log DOR varies with the threshold.

The estimated summary ROC curve can be plotted by computing the expected sensitivity for each value of 1 – specificity across the range of the observed values. The expected sensitivity is given by:

$$\text{Sensitivity} = [1 + e^{-\alpha(1-\beta)}.V^{(1+\beta)(1-\beta)}]^{-1}$$

where V = specificity/(1 – specificity)

List of abbreviations

ANOVA	analysis of variance
AUDIT	Alcohol Use Disorders Identification Test
BMI	body mass index
C	category
CEA	carcinoembryonic antigen
CI	confidence interval
CRD	Centre for Reviews and Dissemination
CRP	C-reactive protein
CT	computed tomography
DARE	Database of Abstracts of Reviews of Effectiveness
DOR	diagnostic odds ratio

continued

List of abbreviations *continued*

DSM-III	Diagnostic and Statistical Manual of Mental Disorders
DVT	deep venous thrombosis
ECG	electrocardiogram
ELISA	enzyme-linked immunosorbent assay
EP	evoked potential
ESR	erythrocyte sedimentation rate
FNAB	fine-needle aspiration biopsy
I	item within a category
ICC	intraclass correlation coefficient
LR	likelihood ratio
LVH	left ventricular hypertrophy
MDP	technetium-99m methylene diphosphate
MI	myocardial infarction
MRI	magnetic resonance imaging
MS	multiple sclerosis
NBT	nitroblue tetrazolium test
NPV	negative predictive value
ns	not significant
PCR	polymerase chain reaction
PE	pulmonary embolism
PPi	technetium-99m (Sn) pyrophosphate
PPV	positive predictive value
QA	quality assessment
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RCT	randomised controlled trial
RDC	research diagnostic criteria
RDOR	relative diagnostic odds ratio
ROC curve	receiver operating characteristic curve
SAAST	Self-Administered Alcoholism Screening Test
SBP	systolic blood pressure
SPECT	single-photon emission computed tomography
STARD	STAndards for Reporting Diagnostic Accuracy
^{99m}Tc	technetium-99m
TRH-ST	thyrotropin-releasing hormone stimulation test
UTI	urinary tract infection
VA	validity assessment

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices in which case the abbreviation is defined in the figure legend or at the end of the table.

Executive summary

Background

The assessment of the quality of studies included in a systematic review is as important for reviews of studies of diagnostic accuracy as it is for any other type of review. There is currently a lack of a validated tool for the assessment of such studies.

Objectives

This project aims to develop a quality assessment tool which will be used in systematic reviews to assess the quality of primary studies of diagnostic accuracy.

Methods

Three systematic reviews were conducted to provide an evidence base for the development of the quality assessment tool. The methodological literature on diagnostic test assessment was reviewed to identify potential sources of bias. Systematic reviews of diagnostic tests that used any form of quality assessment were examined to identify how quality was incorporated. Lastly, a review of existing quality assessment tools was conducted to ascertain what methods exist for assessing the quality of diagnostic studies, and on what evidence they are based. Literature searches were used to identify studies for each of the reviews. Systematic inclusion criteria were applied; studies were selected for relevance and inclusion by one reviewer and checked by a second. Data for each of the reviews were extracted into an Access database by one reviewer and checked by a second. All discrepancies were resolved by discussion or through consultation with a third reviewer when agreement could not be reached. A narrative synthesis is presented for each of the reviews.

A Delphi procedure was used to develop the quality assessment tool. The information provided by the reviews was incorporated into this. A panel of nine experts in the area of diagnostic accuracy studies took part in the Delphi procedure. In the first round members were asked to indicate which of the items on the initial list of items (provided by the results of the reviews) should be included in the tool. Items for which there were high levels of agreement were selected for inclusion/exclusion in the tool; items for which there was disagreement were rated again as part of the next round. Panel members were also asked to make comments and to suggest rephrasings of the items or additional items if appropriate. During subsequent rounds the results of previous rounds were fed back to panel members and they were asked to rerate the items based on the results of the previous rounds. The procedure was continued until agreement was reached on which items were to be included in the quality assessment tools. Panel members were also asked to provide feedback on various other items such as the proposed scoring method, whether they endorsed the procedure, whether they had used the evidence provided to them, and whether they would like to see the development of additional topic and design specific items.

The Delphi procedure produced the quality assessment tool, named QUADAS. A background document was produced which gives details on what is meant by each item included in the tool and how each of the items should be scored.

Work to validate the tool will continue beyond the scope of this project. The validation process will include the piloting of the tool on a small sample of published studies, assessment of the consistency and reliability of the tool, piloting the tool in a number of diagnostic reviews, and using a regression analysis to investigate associations between study characteristics and estimates of diagnostic accuracy in primary studies, as combined in existing systematic reviews.

Results

The reviews produced a list of 28 possible items for inclusion in the quality assessment tool. The first review found that the sources of bias supported by the most empirical evidence were variation by clinical and demographic subgroups, disease prevalence/severity, partial verification bias, clinical review bias and observer/instrument variation. There was also some evidence of bias for

the effects of distorted selection of participants, absent or inappropriate reference standard, differential verification bias and review bias. The evidence for the effects of other sources of bias was insufficient to draw conclusions regarding the effects, if any, of these biases. The third review found that only one item, the avoidance of review bias, was included in more than 75% of tools. A further four items were each included in 50–75% of tools: spectrum composition, population recruitment, absent or inappropriate reference standard and verification bias. Other items were included in less than 50% of tools.

The second review found that the quality assessment tool needs to have the potential to be discussed narratively, reported in a tabular summary, used as recommendations for future research, used to conduct sensitivity or regression analyses and used as criteria for inclusion in the review or a primary analysis. The resulting implication for the development of the tool is that some distinction needs to be made between high- and low-quality studies. It was decided that component analysis is the best approach to incorporate quality into systematic reviews of diagnostic studies. The quality tool was developed taking this into consideration.

The Delphi procedure consisted of four rounds, after which agreement was reached on the items to be included in QUADAS. The final tool included 14 items:

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 classify the target condition correctly?
4. Is the 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?

Conclusions

This project produced an evidence-based quality assessment tool to be used in systematic reviews of diagnostic accuracy studies. Through the various stages of the project the current lack of such a tool and the need for a systematically developed validated tool were demonstrated. Further work to validate the tool continues beyond the scope of this project. The further development of the tool by the addition of design- and topic-specific criteria is proposed.

Chapter 1
Background

Diagnosis can be established either by clinical examination and history taking or by using additional data or information from, for example, a clinical laboratory, radiology or pathology service. Diagnosis can inform or direct therapy decisions and often provides some indication of patient prognosis.

Evaluation methods for diagnostic tests

Studies of diagnostic tests have commonly addressed one of two main objectives, and the chosen study methodology is likely to reflect this. The first, and traditionally the most common aim of diagnostic test evaluation, is to establish the diagnostic accuracy of the test. This is usually done in observational studies reporting test parameters such as sensitivity, specificity, likelihood ratios and the predictive value of the test.[4–6] Measures of sensitivity and specificity can be used together with estimates of the pre-test probability of disease to produce estimates of the post-test probability of disease. For tests with multiple cut-off values, a receiver operating characteristic (ROC) curve can be constructed to assess diagnostic performance. The majority of published studies focus on diagnostic accuracy.

The second objective of diagnostic research is to evaluate the impact of one or more diagnostic strategies on therapy decisions and/or patient outcomes. This trend is in response to the recognition that increasing diagnostic accuracy is of little use if there is no resulting change or improvement in patient care. Such studies are becoming more common, and tend to be either randomised controlled trials (RCTs) or non-experimental comparative studies, both of which are valid under given circumstances. These study designs may also be used in combination, for example, in the comparison of two tests, all patients could receive both tests in (randomised) sequence, with the second test performed without knowledge of the results of the first. This would produce three groups of patients: a group of patients who test positive on both tests who should then all receive treatment; a group of patients who test negative on both tests who then require no treatment; and a group of patients who test positive on one test and negative on the other. This latter group could then be randomised to undergo one therapeutic approach or another.

Other methods of evaluating and combining diagnostic studies are also being developed in order to evaluate diagnostic tests in relation to their clinical or therapeutic impact.[7–10] Decision analytic methods have also been proposed as a means of establishing the impact of a diagnostic technology on patient outcomes and costs.[11,12]

Systematic reviews of diagnostic tests

There is a growing interest in the systematic review and quantitative synthesis of research evaluating diagnostic tests; several reviews have previously been funded by the HTA programme[13–19] and more are ongoing.[20–25] A crucial step in the process of reviewing is the critical appraisal of the quality and results of the primary studies; however, no validated method of assessing study quality is currently available.

Quality of diagnostic studies

The quality of any study can be considered in terms of internal validity, external validity, and the quality of data analysis and reporting. This project will focus on issues of both internal and external validity, although other quality issues will also be considered.

Internal validity can be defined as the degree to which estimates of diagnostic accuracy produced in a study have not been biased as a result of study design, conduct, analysis or presentation (e.g. sample selection, problems with the reference standard and non-independent assessment).

External validity concerns the degree to which the results of a study can be applied to patients in

practice, and is affected by factors such as spectrum of disease or non-disease, setting, other patient characteristics, how the diagnostic test was conducted, the threshold (or cut-off point) used and the reproducibility of the test.

Several methodological reviews have evaluated the quality of reporting of diagnostic studies, the most recent of which found that only one out of seven quality standards, the avoidance of verification bias, was fulfilled by more than 50% of the 112 eligible studies retrieved,[26] although this was found to be a significant improvement on the findings of similar earlier studies.[27–30]

Poor methodological quality has been empirically proven to affect the results of controlled trials and meta-analyses of intervention or treatment studies,[31,32] and this has also been shown for meta-analyses of diagnostic studies.[33] It is generally recognised that study quality should be assessed in any attempt to use results of published studies of diagnostic evaluation.[2,6,34–36] Statistical methods have been developed to account to some degree for, for example, verification bias,[37] as well as to evaluate tests for which there is no (or only an imperfect) reference standard.[38]

Quality assessment tools to assess diagnostic studies

Diagnostic studies have several unique features in terms of quality, for example verification and spectrum bias, which are not addressed by the traditional approach to evaluating controlled trials (which has focused on randomisation, allocation concealment and blinded outcome assessment).

No validated quality assessment tool for diagnostic studies is currently available; however, several researchers have provided guidance for the evaluation of diagnostic studies and suggested criteria necessary for good quality studies.[4–6,39–44] Unfortunately, as Mulrow and co-workers have pointed out, "these criteria have not been uniform … have not been well described and quantitative techniques for assessing quality have not been incorporated".[45] Mulrow and colleagues developed a formal checklist for assessing the quality of a diagnostic test evaluation using expert consensus techniques.[45] Other checklists include those developed for the *JAMA* 'Users' Guide' series,[35,36] the Cochrane Collaboration[46] and others.[5,47–49] Some existing tools are relatively generic, whereas others are more topic specific and they have been used for varying purposes.

There are three main methods of assessing the quality of a study: individual items, checklists, and levels of evidence and quality scores. The best approach for the assessment of diagnostic studies has not yet been identified, although the checklist approach has been recommended because "different design flaws are likely to cause different biases".[34] This is also an argument against using quality scores as it is almost impossible to determine objectively the weighting that should be assigned to each item to produce an overall score.

Assessment tool validation

A method of assessing the quality of a diagnostic study can itself be considered a diagnostic test, which aims to distinguish good-quality from poor-quality studies. Therefore, any such method should be validated according to the principles of establishing the usefulness of any diagnostic test. No existing method to assess the quality of diagnostic studies has been satisfactorily validated, and although the face and content validity of some existing methods seems reasonable, there is a need for empirical evidence of construct validity.

It is difficult, however, to prove the validity of a quality assessment tool owing to the lack of a reference standard against which a new instrument can be measured. In the absence of such a reference standard, an approach that has been used is that utilised by Schulz and co-workers,[32] who assessed the methodological quality of 250 controlled trials from 33 meta-analyses and then used multiple logistic regression models to examine any association between quality assessment and estimated treatment effects. A similar approach has also been followed in the context of diagnostic studies by Lijmer and colleagues.[33]

Using quality assessment in diagnostic test reviews

There are several ways of incorporating the results of quality assessment into a systematic review.[50]

- It may be used as a means of including or excluding studies, where those that do not meet predetermined methodological standards are excluded.

- Studies may be graded according to the quality assessment results, and only those of higher quality included in the primary analysis.
- Studies may be graded according to quality and the results used in sensitivity analysis conducted to examine whether estimates of diagnostic accuracy or other relevant outcomes are associated with the methodological quality of the studies.
- Meta-regression can be used to investigate the effect of either a combined quality score or of various quality components on the pooled estimate.
- A quality assessment score may be produced and used to weight the results of a meta-analysis of individual studies.
- A probability model may be used to try to adjust for quality in forming a summary and making inferences.[51]
- Quality assessment can be used to highlight areas of quality poorly addressed by the studies included in the review, and use this as recommendations for future research.

STAndards for Reporting Diagnostic Accuracy (STARD) project

The STARD project was set up with the aim of formulating an evidence-based statement, including a checklist of items to be described, to improve the quality of the reporting of studies of diagnostic accuracy. There is some overlap between the STARD project and this review: both projects aim to look at sources of variation and bias in diagnostic test evaluations using an evidence-based approach. The projects differ in the aims and the tools that they are developing. The STARD tool aims to provide a single general checklist to act as a guideline for the reporting of all diagnostic research. This project aims to develop a quality assessment tool, which will be used in systematic reviews to assess the quality of primary studies. It is envisaged that this tool will have both a general section relevant to all reports of diagnostic tests and topic-specific sections. For the current project the researchers collaborated with the team at the Academic Medical Centre at the University of Amsterdam who are coordinating the STARD project.

Chapter 2

Research questions

This project had the following aims.

- **Objective 1**: Review the literature on the concepts underlying diagnostic research and identify the main factors that can bias the results of diagnostic studies.
- **Objective 2**: Examine how quality assessment has been used in systematic reviews.
- **Objective 3**: Examine existing methods or assessment tools that have been used to assess the quality of diagnostic research, and any evidence on which they are based.
- **Objective 4**: Develop a new evidence-based assessment tool for the quality assessment of diagnostic studies.

Diagnostic studies are here defined as those studies that examine any procedure, or test, that tries to confirm or identify the presence or absence of a target condition in humans or animals. This includes screening (which can be seen either as a means of identifying high-risk groups, or of identifying disease at an earlier stage), taking patient history, physical examination, biochemical measurements, imaging and invasive procedures, and can also include the evaluation of intermediate outcomes, measures of severity, or 'questionnaire scales', for example, to aid in the diagnosis of mental illness. Such a broad definition of diagnosis is being adopted as similar quality issues should be considered regardless of the procedure or test being evaluated.

It is likely, therefore, that a single assessment tool will not suffice for all circumstances. Instead, the aim of this project is to develop a generic component, which will apply to all or most scenarios. Additional topic-specific elements will be developed beyond the scope of this project. It is anticipated that these will apply to the following contexts:

- screening
- taking patient history/physical examination
- biochemical measurements
- imaging
- invasive procedures
- questionnaire scales.

Chapter 3
Approach

A series of three systematic reviews of the methodological literature were undertaken from which criteria for assessing diagnostic test evaluations were developed.

The methodological literature on diagnostic test assessment was reviewed to identify potential sources of bias. Systematic reviews of diagnostic tests that have used any form of quality assessment were examined to identify how quality was incorporated. Lastly, a review of existing quality assessment tools was conducted to ascertain what methods exist for assessing the quality of diagnostic studies, and on what evidence they are based.

The key sources of bias in diagnostic test assessment were identified from these systematic reviews, and a new assessment tool was developed aimed at addressing these sources of bias.

The final stage in the development of the tool is the validation of the tool. This stage is beyond the scope of this project and will continue after the completion of this project. The stage of developing topic and design specific items will also take place after the completion of the project. The stages involved in this project are illustrated in *Figure 1*.

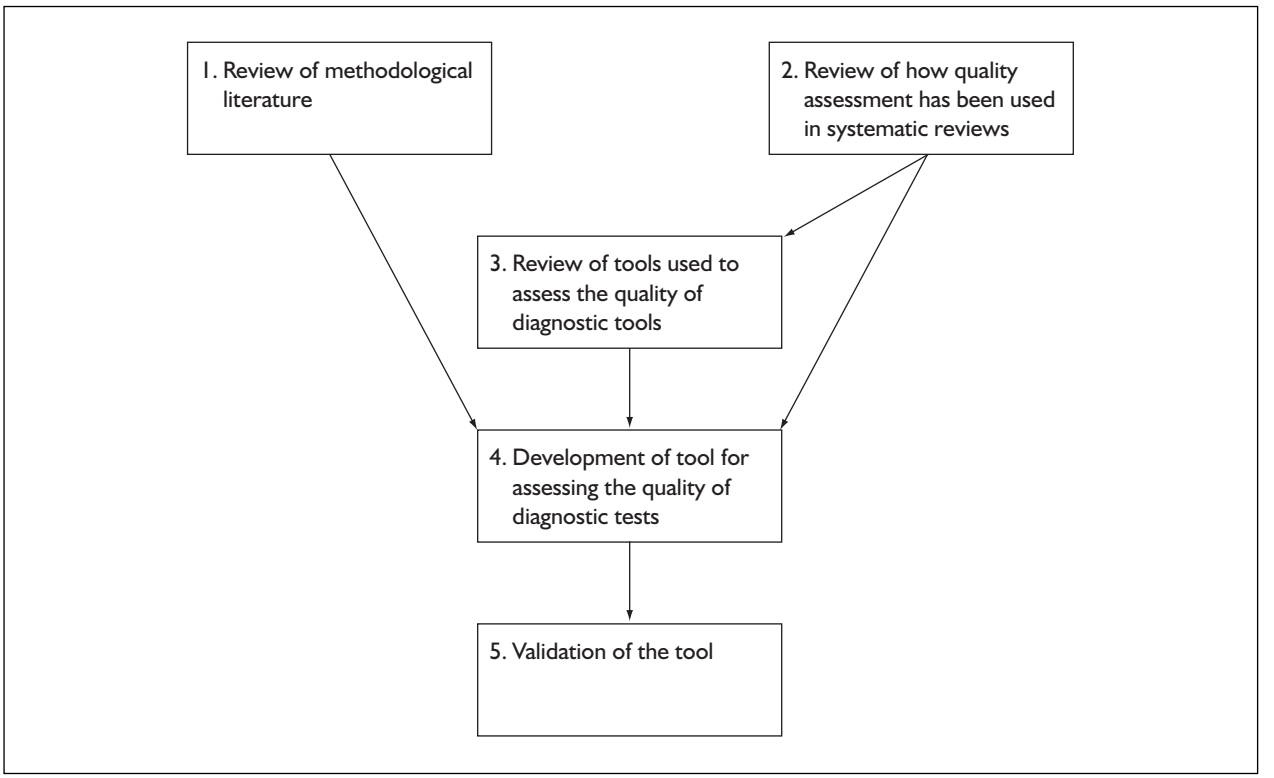

FIGURE 1 *Review flow diagram*

Chapter 4
General methods

An advisory panel was established and invited to offer comment on the protocol and draft report. The systematic reviews (objectives 1, 2 and 3) were undertaken in accordance with the Centre for Reviews and Dissemination (CRD) guidelines for undertaking systematic reviews, with adaptations where necessary.[52]

Classification of bias

The different types of bias that may affect diagnostic test evaluations were grouped into three sections for the analysis of data for objectives 1 and 3, with an additional category for objective 3 (*Table 1*).

Chapters 5 and 7 give detailed justification of these lists.

Spectrum composition

Variation by clinical and demographic subgroups (spectrum composition)
Measures of diagnostic accuracy can vary substantially when applied to populations with different demographic and clinical features; this is known as spectrum bias. Reported estimates of diagnostic accuracy may have limited clinical applicability (generalisability) if the spectrum of tested patients is not adequately described. The spectrum of patients refers not only to the severity of the underlying target condition, but also to demographic features and to the presence of differential diagnosis and/or co-morbidity. It is therefore important that diagnostic test evaluations include an appropriate spectrum of patients for the test under investigation and that a clear definition of the characteristics of the included patients is provided. To optimise generalisability, the test should be evaluated in the population in which it will be used in practice. This is discussed in more detail below. Variation in estimates of test performance related to spectrum composition does not necessarily affect the internal validity of the study, but limits the clinical applicability of study results. The word bias is therefore something of a misnomer.

Inclusion criteria (objective 3 only)
This refers to whether studies have provided a clear definition of the criteria used as criteria for entry into the study. It is related to all of the other three items included under the subheading 'spectrum composition', as the provision of detailed information regarding the inclusion criteria will allow an assessment of whether an appropriate spectrum of patients has been included in the study. This item is only included for objective 3 as it does not itself lead to variation in estimates of test performance, but may be an important feature for studies to report and has therefore been included in existing checklists.

Distorted selection of participants (population recruitment)
The selection process of patients within a diagnostic study is a vital step because it determines the composition of the study population (spectrum). Knowledge about the composition of the study population is essential in the judgement of applicability, as measures of diagnostic accuracy may vary with clinical and other features. Patient selection in diagnostic studies can have different starting points. The ideal form of patient inclusion is the one that includes all patients suspected of the target condition within a specific period. This is known as a consecutive series. This selection procedure is frequently used in diagnostic cohort designs. In the cohort design, all patients usually receive the index test before the reference standard. In other designs, in which the flow is reversed and patients receive the reference standard before the index test, sampling of patients becomes more complex to unravel. If the flow is reversed and the selection is based on an already known disease state, this is similar to a conventional case–control study and is often called a diagnostic case–control. In this design, the enrolment of cases is based on positive test results on the reference standard, whereas the controls may be selected in different ways. The control group can consist of healthy controls or a convenience sample (e.g. all patients present at the department). The control group can also consist of patients with disorders classified within the differential diagnosis of the target disorder. Other designs start their inclusion with patients who did receive the reference standard or the

TABLE 1 *Classification of bias*

Review of methodological literature (Chapter 5)	Review of tools used to assess the quality of diagnostic tests (Chapter 7)
Spectrum composition (patients)	
Variation by clinical and demographic subgroups	Spectrum composition
	Inclusion criteria
Distorted selection of participants	Population recruitment
Disease prevalence/severity	Disease prevalence/severity
Index test and reference standard	
Selection and execution	
Absent or inappropriate reference standard	Absent or inappropriate reference standard
Change in technology of index test	Change in technology of index test
Disease progression bias	Disease progression bias
Difference in test protocol	Test execution
	Reference execution
Partial verification bias	Verification bias
Differential verification bias	
Incorporation bias	Incorporation bias
	Normal defined
Treatment paradox	Treatment paradox
Interpretation	
Review bias	Review bias
Clinical review bias	Clinical review bias
Observer/instrument variation	Observer/instrument variation
Analysis	
	Appropriate results
Precision (sample size, variation by chance)	Precision (sample size, variation by chance)
Inappropriate handling of uninterpretable/indeterminate/ intermediate test results	Inappropriate handling of uninterpretable/indeterminate/ intermediate test results
Post hoc choice of threshold value	Post hoc choice of threshold value
Dropouts	Dropouts
	Subgroups
	Data table
	Utility of test
Research planning	
	Sample size
	Objectives
	Protocol

index test or both. The selection process is more difficult to reconstruct when existing data sources (e.g. hospital records, registers) have been used.[53] Theoretically, the method of patient inclusion would be expected to affect estimates of test performance, mainly through the impact on spectrum composition. If consecutive patients are enrolled then the spectrum composition can be very different compared with a study in which a case–control design is used.

Disease prevalence/severity

The setting of a diagnostic accuracy study and the associated referral pattern of patients may affect estimates of diagnostic accuracy. The prevalence of the target condition varies according to setting and referral pattern, and altering the prevalence may affect some measures of diagnostic accuracy. Negative and positive predictive values are directly affected by changes in prevalence, assuming constant sensitivity and specificity (Bayes' theorem). Although there is no mathematical relationship between sensitivity and specificity on the one hand and prevalence on the other, the mechanism responsible for the change in prevalence may also act on sensitivity or specificity. One example would be a referral mechanism leading to both an increase in prevalence and a higher proportion of more severe cases after referral. After referral both prevalence and sensitivity would be higher.[54] Context bias, the tendency for interpreters to consider test results more frequently abnormal in settings with higher disease prevalence, may be a problem in diagnostic test evaluations.[55] A test would be expected to perform better in populations with

more severe disease as it is generally easier to detect than mild disease. After referral, in general, the prevalence will be higher and there tend to be more severe cases. This link between severity and prevalence could explain why sensitivity is often higher in situations with higher prevalence.

Index test and reference standard

Selection and execution

Absent or inappropriate reference standard

The reference standard is the test used to measure the presence or absence of the target condition. To assess the diagnostic accuracy of the index test, its results are compared with the results of the reference standard; subsequently, indicators of diagnostic accuracy can be calculated. The reference standard is therefore an important determinant of the diagnostic accuracy of a test. The reference standard may be obtained in many ways, including laboratory tests, imaging tests, function tests and pathology, but also clinical follow-up of participants. These tests range from radiography to autopsies. The decision of which reference standard to use depends on the definition of the target condition and the purpose of the study. If no single reference standard is available, the most likely state of the patients can be derived from careful clinical follow-up or a consensus between observers,[56] or modelled from results of two or more index tests.[57–59] The reference standard is a proxy for the target condition and therefore often not perfect. Reference standard error bias occurs when errors of imperfect reference standard(s) bias the measurement of diagnostic accuracy of the index test.[57,60] If there are any disagreements between the reference standard and the experimental test then it is assumed that the experimental test is incorrect.[1] From a theoretical point of view the choice of an appropriate reference standard would appear to be very important. Estimates of test performance are based on the assumption that the test is being compared to a reference standard that is 100% sensitive and specific. If this is not the case then it may be that the index test classifies results correctly that have been incorrectly classified by the reference standard. This would therefore be expected to give an underestimation of the performance of the index test.

Change in technology of index test

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.[61] This may be an important feature to take into consideration when looking at studies conducted at different points in time. Differences in technology may also impair the comparability of studies conducted independently at different centres, leading to similar problems with estimates of test performance.

Disease progression bias

Time delay may lead to disease progression bias. This occurs when there is an abnormally long period between the performance of the test under evaluation and the confirmation of the diagnosis with the reference standard, so that the disease is at a more advanced stage when the reference standard is performed.[62] The converse may occur if the reference standard is performed first. Short-term fluctuations may also be important; for example, diurnal variation in cortisol levels or short-term fluctuation in blood pressure. Whether time delay is a problem for a diagnostic test evaluation will depend on the particular test context.

Difference in test protocol (index test and reference standard execution)

A sufficient description of the execution of index test and reference standard is important for two reasons. First, variation in measures of diagnostic accuracy can sometimes be traced back to differences in the execution of index test/reference standards. Second, a clear and detailed description (or references) is needed to implement a certain test in another setting. This may be a challenge, especially if translation is involved. If tests are executed in different ways then this would be expected to impact on test performance. The extent to which this would be expected to affect results would depend on the type of test being investigated.

Verification bias (objective 3 only)

Verification bias occurs when not all of the study group receives definitive confirmation of the diagnosis with the same reference standard.[62] If the selection to receive the index test is a random subgroup of the positive (or negative) patients the overall diagnostic performance of the test is, in theory, unchanged. In most cases, however, this selection is not random, leading to biased estimates of the overall diagnostic accuracy. Factors other than index test results, such as age and gender (concomitant information factors[63]), can influence the decision to order the reference standard. This may lead to secondary selection bias.[64] For objective 1 only, a distinction was made between partial and differential verification bias.

Partial verification bias (objective 1 only)
Partial verification bias occurs when only a selected sample is verified by the reference standard. Theoretically, partial verification bias would be expected to have a significant impact on estimates of test performance. If only a selected sample of patients undergoes the reference standard, usually because they tested positive on the index test, then this would be expected to affect test performance. If not all patients with negative test results are verified by the reference standard, and those who do not receive verification are omitted from the 2×2 table, both sensitivity and specificity may be affected. In this situation patients with false-negative test results will stay undetected, which inflates sensitivity. Conversely, the exclusion of patients with true-negative test results will artificially decrease specificity.

Differential verification bias
Differential verification or reference standard bias occurs when part of the index test results is verified by a different reference standard. This is especially a problem if these reference standards differ in their definition of the target condition; for example, histopathology of the appendix and natural history for the detection of appendicitis.[33,65] This usually occurs when patients testing positive on the index test receive a more accurate, often invasive, reference standard test than those with a negative test result. The link (correlation) between a particular (negative) test result and being verified by a less accurate reference standard will affect measures of test accuracy in a similar way as in partial verification, but less seriously.

Incorporation bias
When the result of the index test is used in establishing the final diagnosis incorporation bias may occur.[66–68] This incorporation will probably increase the amount of agreement between index test results and the outcome of the reference standard, and hence overestimate the various measures of diagnostic accuracy.

Normal defined (objective 3 only)
Whether a definition is provided of which test results would be considered 'normal' and which would be considered 'abnormal' may be an important feature of a diagnostic test evaluation. The provision of such a definition will not directly bias the estimation of test performance and so it is not included for objective 1. However, it may be an important feature to be aware of when looking at diagnostic test evaluations and so this is often included in quality assessment tools for studies of diagnostic accuracy.

Treatment paradox
If treatment is started based on the knowledge of the results of the index test, and the reference standard is applied after treatment has started, then treatment paradox may bias estimates of test performance. Whether or not this form of bias will affect test results will be related to the condition being investigated. For conditions that will respond quickly to treatment, such as bacterial infections treated with antibiotics, this is more likely to be a problem than for chronic conditions that are more difficult to treat.

Interpretation

Review bias
Review bias occurs 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.[69] Test review bias occurs when results of the reference standard are known while interpreting the index test.[70] Comparator review bias occurs when a diagnostic study involves the comparison of two or more diagnostic tests with the reference standard, and the results of any of the tests are known when interpreting results.[62] Review bias is in concept similar to the effects of non-blinded measurement of outcomes in intervention studies. The extent to which this may affect test results will be related to the degree of subjectiveness in the interpretation of the test result. The more subjective the interpretation the more likely that the interpreter can be influenced by the results of the index test in interpreting the reference standard, and vice versa.

Clinical review bias
The availability of information on clinical data, such as age, gender and symptoms, during interpretation of test results may affect estimates of test performance.[62] The knowledge of such factors can influence the diagnostic test result if the test involves an interpretative component.[61]

Observer/instrument variation
The reproducibility (also referred to as reliability) of test results is one of the determinants of diagnostic accuracy of an index test. Because of variation in laboratory procedures or observers, a test may not consistently yield the same result when repeated. If the reproducibility of a test is

not perfect, there will always be some false-positive and false-negative results. Hence, the accuracy of the test cannot be perfect either.

Observer variability can arise during description, when the observed entity is converted into data, or during classification, when the data are converted into diagnostic or other stipulated categories. In two or more observations of the same entity, intraobserver variability arises when the same person obtains different results, and interobserver variability, when two or more people disagree. Instrument variability concerns the amount of variation that arises during the operation of devices or systems, such as automated laboratory measurements. This kind of variation is also referred to as analytical methodological variation or analytical noise (error). As with review bias, the extent to which observer variability may affect test results will be strongly related to the subjective component involved in their interpretation.

Analysis

Appropriate results (objective 3 only)
The presentation of appropriate results, such as sensitivity, specificity, likelihood ratios and predictive values, may be an important indicator of whether a diagnostic test evaluation study has been undertaken appropriately. This will not have any effect on biasing estimates of test performance and so is not included for objective 1. However, it is often included in quality assessment tools and so is included for objective 3.

Precision (sample size, variation by chance)
The final aim of a study on diagnostic accuracy is to produce an expression on how well the test results correspond with the presence or absence of the target condition, as established by the reference standard. These estimates are subject to sampling variation. Therefore, authors should present a confidence interval around an estimate of diagnostic accuracy, which quantifies the amount of statistical uncertainty (degree of precision) around the observed value.[71] It will show the reader the range of likely values around an estimate of diagnostic accuracy that is compatible with the observed data.

Inappropriate handling of uninterpretable/ indeterminate/intermediate results
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. Whether bias will arise depends on the possible correlation between uninterpretable test results and the true disease status. The direction and size of the bias result from a complicated interaction among the indeterminate test result, the 'true' underlying index test result and the disease status.

Post hoc choice of threshold value
The selection of the threshold value for the index test being evaluated during data analysis by selecting the cut-off point to maximise 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 will be lower, even if the study consists of patients from the same population (spectrum). This optimism is a well-known statistical phenomenon in the literature of prognostic modelling.[72]

Dropouts
This occurs when patients withdraw from the study before the results of both the index test and reference standard are known. If patients lost to follow-up differ systematically from those who remain, for whatever reason, then estimates of test performance may be biased.[62]

Subgroups (objective 3 only)
The presentation of appropriate subgroup analyses will not affect estimates of test performance, but may provide additional relevant information about the performance of a test. For this reason, this item is included for objective 3 but not for objective 1.

Data table (objective 3 only)
Similarly, the presentation of the results of the diagnostic test evaluation in a 2×2 table, or the presentation of sufficient information to calculate 2×2 table data, will not affect diagnostic test performance, but may provide an indication of the overall quality of the study. It is therefore only included for objective 3.

Utility of test (objective 3 only)
The utility of the test refers to how useful the test will be in practice. This will not have any effect on the diagnostic test performance, but may provide an indication of the clinical applicability of the test.

Research planning (objective 3 only)

This classification is only included for objective 3. Items relating to research planning will not have

any direct effect on estimates of diagnostic test performance, but may give an indication as to whether the study has been conducted appropriately and hence the overall quality of the study.

Sample size
It is important that appropriate numbers of diseased and non-diseased participants are included in the study so that the confidence intervals for both sensitivity and specificity will not be too wide.

Objectives
The study objectives should be clearly defined a priori, so that the aim of the study is clear and has been planned appropriately.

Protocol
Protocols are important in showing that a study has been planned and thought about before being started.

Chapter 5

Objective 1: Review the literature on the concepts underlying diagnostic research and identify the main factors that can bias the results of diagnostic studies

Measures of test accuracy are not fixed constants. Even if tests are evaluated in a well-designed study, including features such as consecutive patients, complete verification by appropriate reference standard, independent blind assessment of index tests and reference standard, diagnostic performance may vary from one study to another. Variability due to chance is not the only explanation for this. Other sources of variation include:[73]

- differences in population
- differences in definition of target condition
- differences in tests (technical characteristics, differences in execution or reading of tests)
- different criteria of positivity (threshold values).

These sources of variation have the potential to bias or modify results if they become incorporated into the design of a study. For example, using inappropriate inclusion and/or exclusion criteria will lead to a biased sample of patients with and without the target condition. This distorted study population is not representative for any setting and spectrum bias will ensue. Variation in measures of diagnostic accuracy may come from artefactual differences (e.g. different design features of studies) or true differences (e.g. different test types or different spectrum of disease). True variation can only be discussed after artefactual differences have been addressed. This chapter will focus on the various factors that can lead to variation and/or bias in diagnostic studies. The aim is to provide an indication of those factors that are likely to have an important effect on test results and hence which factors should be included in the quality assessment tool being developed as part of this project. For this purpose a systematic review of all studies that looked at the effects of bias on estimates of test performance was conducted.

Methods

Inclusion criteria

All studies, of any design, with the main objective of addressing bias or variation in diagnostic tests were included in the review. 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 provide an estimation of the extent to which a particular bias may distort these estimates. Inclusion was assessed by one reviewer and checked by a second. Discrepancies were resolved through discussion.

Literature searches

A database of published and unpublished methodological literature was assembled from systematic literature searches, using electronic sources, handsearching and consultation with methodological experts.

Searches of databases including MEDLINE, EMBASE, BIOSIS and the methodological databases of both CRD and the Cochrane Collaboration were performed to identify methodological literature. Citation searches of key papers were undertaken. Full details of the search strategy are provided in Appendix 1. No language restrictions were applied.

Literature in this area is poorly indexed and difficult to locate; therefore, contact with relevant methodological experts was a key source of identifying additional literature. Groups at the Universities of Amsterdam, Maastricht, Leuven and Ottawa who are also conducting work in this field were contacted.

Data extraction

Data were extracted on the following:

- author

- year of publication
- study design (diagnostic cohort, diagnostic case–control, meta-analysis, computer model, statistical model)
- study objective
- type of analysis (statistical or narrative)
- sources of bias addressed (see list below)
- evidence provided on sources of bias addressed (separately for theoretical and empirical).

Data were extracted by one reviewer and checked by a second reviewer. Discrepancies were resolved by consensus or consultation with a third reviewer.

Data synthesis

A narrative synthesis is presented. Results were stratified according to category of bias or variation, classified as follows:

Spectrum composition
- Variation by clinical and demographic subgroups
- Distorted selection of participants
- Disease prevalence/severity

Index test and reference standard

Selection and execution
- Absent or inappropriate reference standard
- Change in technology of index test
- Disease progression bias
- Difference in test protocol
- Partial verification bias
- Differential verification bias
- Incorporation bias
- Treatment paradox

Interpretation
- Review bias
- Clinical review bias
- Observer/instrument variation

Analysis
- Precision (sample size, variation by chance)
- Inappropriate handling of uninterpretable/ indeterminate/intermediate test results
- Post hoc choice of threshold value
- Dropouts.

A description of what is meant by each category of bias is presented in Chapter 4. The evidence relating to each source of bias or variation is presented within each category. Studies were grouped into 'real-life' and 'numerical' studies. Real-life studies were those that used actual data from one or more clinical studies to demonstrate the effect of a particular study feature. These were classified further as diagnostic accuracy studies or experimental studies. Experimental studies were those that were specifically designed to test a hypothesis about the effect of a certain feature; for example, rereading sets of X-rays while controlling (manipulating) the overall prevalence of abnormalities. Diagnostic accuracy studies were those that used either a diagnostic cohort or case–control design. These were classified further according to whether the data collection in the study was retrospective or prospective. Numerical studies were those that used statistical or computer models to simulate how certain types of biases may affect estimates of diagnostic test performance. The results of real-life studies are considered more informative than those of modelling studies as these are based on actual data rather than on models. The results of real-life studies were considered as 'empirical evidence of bias', whereas those of numerical studies were considered as 'theoretical evidence of bias'. The results of studies that provide theoretical evidence of bias are presented in italics so that a distinction can be drawn between those studies providing empirical and theoretical evidence of bias. The effects of each source of bias or variation on estimates of test performance found by each study are summarised in a table and discussed narratively. A summary table is provided showing the number of studies that found empirical evidence of bias, theoretical evidence of bias and no evidence of bias, separately for each source of bias or variation.

Results

The nature of the evidence

The literature searches identified a total of 8663 references. Of these, 569 studies were considered potentially relevant and were assessed for inclusion, and 55 met inclusion criteria. The year of publication of the included studies ranged from 1963 to 2000. Individual study results are presented in Appendix 2. A narrative analysis was provided in three studies[66,74,75] and a statistical analysis in the remaining 52 studies. Forty-seven studies provided empirical evidence of bias and eight provided theoretical evidence (three studies provided both forms of evidence). A diagnostic accuracy design was used in 22 studies, of which 13 were prospective and nine retrospective. Nine studies were reviews and 16 studies used an experimental design.

The studies that used a diagnostic accuracy design used various methods to investigate how bias or variation may affect estimates of test performance. All but one of these studies presented a statistical analysis of the effects of bias or variation. The number of sources of bias or variation investigated ranged from one to three, with the majority of studies looking at one source. Ten studies performed subgroup analysis, including regression analysis in several of these, to compare estimates

of test performance across various patient subgroups. Six studies used statistical methods to correct for verification bias and compared the corrected estimates to the actual estimates obtained from the study in which verification bias was acting. In two studies, in which only a sample of patients had originally received the reference standard, the remainder of the patients was also given the reference standard, and estimates of test performance were compared between the two groups. In another study all the patients had received the reference standard, but a simulation was carried out in which only patients who fulfilled certain conditions were assumed to have received the reference standard, and estimates of test performance were compared between the subgroup and the whole sample. In the final study patients who had been included in 'early' evaluations of a specific diagnostic test were compared with those included in 'recent' evaluations to investigate why estimates of test performance had been declining.

Of the studies that conducted a review, two provided a narrative synthesis combining a limited number of studies, three used meta-analysis to compare studies with different characteristics, three used multivariate analysis to investigate the effects of several different potential sources of bias, and one conducted a review of reviews in which regression analyses were conducted to investigate how several different sources of bias affected study results. The number of sources of bias or variation investigated by each review ranged from one to eight.

The experimental studies all used similar designs based on the interpretation of samples for which the true result was known. All studies provided a statistical analysis of the effects of bias. In five studies a number of samples (radiographs, mammograms, etc.) was interpreted by several different observers on one occasion, in ten studies test results were examined by the same set of observers on two separate occasions, usually separated by a washout period of several months. Two of these studies used this design to measure intraobserver variability, one to investigate the effect of different disease prevalences, and seven to investigate whether the provision of clinical history had any effect on test performance. One study grouped observers into three groups; in two of the groups clinical information was provided for half the samples and in the third group no clinical information was provided.

All of the studies that provided theoretical evidence of bias used some form of modelling to estimate the effects of bias on test performance. The range of methods varied from simple if–then models (four studies) to computer simulations (three studies) and Monte Carlo modelling (one study). All of these studies only looked at the effects of one type of bias.

Spectrum composition

Variation by clinical and demographic subgroups

Fifteen studies investigated the effects of variations in clinical and demographic features on test performance (*Table 2*). All but one of these studies found an association between the features investigated and test performance. All of these provided empirical evidence of bias and one study also provided theoretical evidence. Gender was the most commonly investigated variable. Three studies found no association between test performance and gender, nine found significant effects on sensitivity and four found significant effects on specificity. Other variables shown to have significant effects on test performance were age, race and smoking status. Disease-related variables and co-morbidities were found to be associated with both sensitivity and specificity in six studies and with sensitivity in one study. Three of the 15 studies also looked at the description of the study population. One study found that when no description of the study population was provided test performance was overestimated, one found non-significant effects on test performance, and the third found no association between adequate definition of study group and test performance.

Distorted selection of participants

Five studies looked at the effects of distorted selection of participants on test performance (*Table 3*). Three studies provided empirical evidence of bias. Test performance was found to be overestimated if reasons for exclusion commonly used by researchers were applied, and another study found that *in vivo* studies give higher estimates of test performance compared with *in vitro* studies. The third study found that case–control studies overestimate test performance; the same study found that non-consecutive patient enrolment did not affect performance. Two studies found that the avoidance of a limited challenge group did not have significant effects on test performance.

Disease prevalence/severity

Six studies looked at the effect of disease prevalence and three looked at disease severity (*Table 4*). All nine studies found empirical evidence of bias and one also found theoretical

TABLE 2 *Results of studies that looked at biases associated with variation by clinical and demographic subgroups*

Study details	Effect of bias	Category
Curtin, 1997[76] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	No effect of gender on sensitivity or specificity; weight had strong effect on sensitivity, within weight subgroups gender had an effect, specificity was high in all groups. Sensitivity was higher in heavier participants	Empirical evidence of bias
Detrano, 1988[77] Type of analysis: statistical Study design: review	The proportion of patients with myocardial infarction increased sensitivity. The proportion of men in the study group was related to test sensitivity. Age and use of medications did not affect sensitivity or specificity	Empirical evidence of bias
	Adequate definition of the study group was not associated with test performance	No evidence of bias
Detrano, 1989[78] Type of analysis: statistical Study design: review	Sensitivity and specificity were associated with various disease-related factors; however, some other factors including the exclusion of women showed no association	Empirical evidence of bias
Detrano, 1988[79] Type of analysis: statistical Study design: review	Sensitivity and specificity were associated with various disease-related factors; however, some other factors including age showed no association with test performance. Sensitivity was significantly associated with gender and previous myocardial infarction in the multivariate analysis	Empirical evidence of bias
	Adequate definition of the study group had non-significant effects on sensitivity and specificity	No evidence of bias
Hlatky, 1984[80] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Various disease-related factors and age and gender were associated with sensitivity; less evidence for association with specificity	Empirical evidence of bias
Levy, 1990[81] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Sensitivity was affected by gender, age, obesity and smoking. There was no significant effect on specificity, which was high in all groups	Empirical evidence of bias
Lijmer, 1999[33] Type of analysis: statistical Study design: review	Diagnostic performance was overestimated when no description of the population under study was provided	Empirical evidence of bias
Melbye, 1993[82] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Specificity increased with increasing prevalence of pneumonia and the likelihood ratio dropped	Empirical evidence of bias
Moons, 1997[83] Type of analysis: statistical	Sensitivity differed according to gender and factors related to severity of disease. Variation over smoking and various disease-related factors was less marked. Specificity differed according to gender, diabetes and disease-related factors	Empirical evidence of bias
Morise, 1994[84] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Sensitivity and specificity were higher in men than in women	Empirical evidence of bias
Morise, 1995[85] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	Sensitivity and specificity were higher in men than in women	Empirical evidence of bias

continued

TABLE 2 *Results of studies that looked at biases associated with variation by clinical and demographic subgroups (cont'd)*

Study details	Effect of bias	Category
Roger, 1997[64] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	After correction for verification bias sensitivity was lower in women than in men	Empirical evidence of bias
Santana-Boado, 1998[86] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	There was no difference in sensitivity and specificity between men and women. Earlier observed differences were due to verification bias	No evidence of bias
Stein, 1993[87] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	Test performance was better for patients without prior cardiopulmonary disease	Empirical evidence of bias
Steinbauer, 1998[88] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Test performance was associated with race and gender; a different test was not affected by these variables	Empirical evidence of bias

TABLE 3 *Results of studies that looked at biases associated with distorted selection of participants*

Study details	Effect of bias	Category
Detrano, 1988[77] Type of analysis: statistical Study design: review	Avoidance of a limited challenge group did not significantly affect test performance	No evidence of bias
Detrano, 1988[79] Type of analysis: statistical Study design: review	Avoidance of a limited challenge group had non-significant effects on test performance	No evidence of bias
Lijmer, 1999[33] Type of analysis: statistical Study design: review	Case–control studies overestimated test performance	Empirical evidence of bias
	Non-consecutive patient enrolment did not affect performance	No evidence of bias
Philbrick, 1982[89] Type of analysis: narrative Study design: diagnostic accuracy study, prospective	Test performance was overestimated if reasons for exclusion commonly used by researchers are applied	Empirical evidence of bias
van Rijkom, 1995[90] Type of analysis: statistical Study design: review	*In vivo* studies gave higher estimates of test performance than *in vitro* studies	Empirical evidence of bias

evidence. The effects of disease prevalence and severity on test performance were mixed. In general, sensitivity was found to increase with increased disease prevalence and specificity was found to decrease. One study found that as disease prevalence increases both sensitivity and specificity increase, one found increased sensitivity but decreased specificity, two found increased sensitivity but no effect on specificity, one found increased sensitivity but did not report on the effects on specificity, and one found decreased specificity but did not report on the effects on sensitivity. All three studies that looked at diseased severity found increased sensitivity with more severe disease, two of these found no effect on specificity and the third did not comment on the effects on specificity.

TABLE 4 *Results of studies that looked at biases associated with disease prevalence/severity*

Study details	Effect of bias	Category
Egglin, 1996[55] Type of analysis: statistical Study design: experimental	Sensitivity increased with increased disease prevalence. Specificity was not affected	Empirical evidence of bias
Lachs, 1992[92] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Sensitivity was higher in the group with the highest pre-test probability of disease. Specificity was lower in the group with the higher pre-test probability of disease	Empirical evidence of bias
Levy, 1990[81] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Sensitivity was affected by severity of disease. There was no significant effect on specificity, which was high in all groups	Empirical evidence of bias
Moons, 1997[83] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Sensitivity differed according to severity of disease. Specificity was not affected	Empirical evidence of bias
O'Connor, 1996[93] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Sensitivity was greatest in groups with a higher pre-test probability of disease. There were no significant differences in specificity	Empirical evidence of bias
Ransohoff, 1978[66] Type of analysis: statistical Study design: review	Sensitivity may be overestimated in a diseased group with extensive disease	Empirical evidence of bias
Rozanski, 1983[94] Type of analysis: statistical Study design: review	Specificity was lower in patients with a higher pre-test probability of disease	Empirical evidence of bias
Taube, 1990[95] Type of analysis: statistical Study design: modelling with an example using diagnostic accuracy design	*Theoretical: the less advanced the disease the lower the sensitivity*. Example supported these results	Empirical and *theoretical* evidence of bias
van der Schouw, 1995[53] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	As selection criteria were widened (and disease prevalence increased) both sensitivity and specificity increased	Empirical evidence of bias

Index test and reference standard

Selection and execution

Absent or inappropriate reference standard

Eight studies looked at reference standard error bias (*Table 5*). Four studies found empirical evidence of bias and four found theoretical evidence. One of the four studies that provided empirical evidence of bias found that weaker validation methods may overestimate test performance, the second found that different reference standards can provide very different estimates of test performance, the third found that studies which used a specific reference standard (tomographic imaging) overestimated test performance compared with other studies, and the fourth found that comparison with a more accurate test was related to sensitivity. 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

TABLE 5 *Results of studies that looked at biases associated with an absent or inappropriate reference standard*

Study details	Effect of bias	Category
Arana, 1990[96] Type of analysis: statistical Study design: review	Different reference standards provided very different estimates of test performance	Empirical evidence of bias
Boyko, 1988[97] Type of analysis: statistical Study design: modelling	*With an imperfect reference standard specificity is best estimated at low disease prevalence and sensitivity at high disease prevalence. Considerable errors in estimates exist, even when the reference standard has close to perfect performance (96% sensitivity and specificity)*	*Theoretical evidence of bias*
De Neef, 1987[98] Type of analysis: statistical Study design: modelling	*When the new test is more accurate than the reference standard test performance is underestimated. Estimates strongly related to disease prevalence*	*Theoretical evidence of bias*
Detrano, 1989[78] Type of analysis: statistical Study design: review	The comparison with another exercise test thought to be superior in accuracy was found to be significantly and independently related to sensitivity	Empirical evidence of bias
Detrano, 1988[79] Type of analysis: statistical Study design: review	Studies that used tomographic imaging as the reference standard overestimated test performance	Empirical evidence of bias
Phelps, 1995[99] Type of analysis: statistical Type of design: modelling	*When diagnostic test errors are statistically independent inaccurate reference standards lead to underestimation of test performance; when they are dependent can lead to overestimation*	*Theoretical evidence of bias*
Thibodeau, 1981[100] Type of analysis: statistical Type of design: modelling	*Studies that compare the index tests to a reference standard which contain errors will underestimate test performance, as long as the diagnostic test is more often positive in the diseased than in the non-diseased. If conditional dependence is present then test performance will be even lower*	*Theoretical evidence of bias*
van Rijkom, 1995[90] Type of analysis: statistical Study design: review	Weaker validation methods may overestimate test performance	Empirical evidence of bias

TABLE 6 *Results of studies that looked at biases associated with a change in the technology of the index test*

Study details	Effect of bias	Category
Detrano, 1988[79] Type of analysis: statistical Study design: experimental	Automation of test improved sensitivity but decreased specificity	Empirical evidence of bias
Froelicher, 1998[101] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	No difference was found between computerised readings and physician readings	No evidence of bias

other two theoretical studies found that test performance is underestimated when the test being evaluated is more accurate than the reference standard.

Change in technology of index test

Two studies were identified which looked at the effects of a change in the technology of the index test on test performance (*Table 6*). One study found that automation of the test procedure improved test sensitivity but decreased specificity, providing empirical evidence of bias; the other found no effect on test performance.

Disease progression bias

Only one study was identified which looked at the effects of disease progression bias on test

TABLE 7 *Results of studies that looked at biases associated with disease progression bias*

Study details	Effect of bias	Category
Detrano, 1988[79] Type of analysis: statistical Study design: review	The maximum interval between the index and reference standard was not associated with test performance	No evidence of bias

TABLE 8 *Results of studies that looked at biases associated with reporting of execution of index and reference standards*

Study details	Effect of bias	Category
Detrano, 1988[79] Type of analysis: statistical Study design: review	Exercise protocol was not significantly related to test performance	No evidence of bias
Lijmer, 1999[33] Type of analysis: statistical Study design: review	Failure to describe index test execution overestimated test performance, failure to describe reference standard execution decreased test performance	Empirical evidence of bias

performance (*Table 7*). This study found no evidence of bias.

Difference in test protocol
Only two studies looked at the effects of execution of tests (*Table 8*). One study found that failure to describe the index and reference standard execution biases estimation of test performance and provided empirical evidence of bias. The other study found no effect of exercise protocol on test performance.

Partial verification bias
Twenty-two studies investigated the effects of partial verification bias (*Table 9*). Three studies found no evidence of bias, two provided theoretical evidence of bias, one provided both theoretical and empirical evidence of bias, and the remaining 16 studies provided empirical evidence of bias. The effects of verification bias are, however, unclear. Of the two studies that provided theoretical evidence of bias, one stated that verification bias biases estimates of test performance and the other found that verification bias increases sensitivity and decreases specificity. Two of the studies that provided empirical evidence of bias reported that test performance was increased, but the effects on sensitivity and specificity were not reported. Two studies reported that estimates of test performance were biased but provided no further information. Eight studies found that sensitivity was increased and specificity decreased in the presence of verification bias, one study found that both sensitivity and specificity were increased, two found that specificity was increased and two found that specificity was decreased.

Differential verification bias
Only two studies looked at differential verification bias (*Table 10*). Both of these found that differential verification bias may be associated with test performance and provided empirical evidence of bias.

Incorporation bias
No studies were identified which provided evidence of the effect of incorporation bias.

Treatment paradox
No studies were identified which provided evidence of the effect of treatment paradox.

Interpretation
Review bias (test and diagnostic)
Five studies investigated review bias; four studies looked at both diagnostic and test review bias and one looked only at diagnostic review bias (*Table 11*). Two studies found that blinding reduces the concordance between test results, one study found that failure to avoid review bias may overestimate sensitivity and specificity, one found that diagnostic review bias can overestimate test performance and the last found that blinding was not associated with test performance. Four studies provided empirical evidence of bias and one study found no evidence of bias.

TABLE 9 *Results of studies that looked at biases associated with verification bias*

Study details	Effect of bias	Category
Bowler, 1998[102] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	The study shows considerable scope for verification bias	Empirical evidence of bias
Cecil, 1996[103] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	Partial verification bias overestimated sensitivity and underestimated specificity	Empirical evidence of bias
Detrano, 1988[77] Type of analysis: statistical Study design: review	Verification bias overestimated specificity	Empirical evidence of bias
Detrano, 1989[78] Type of analysis: statistical Study design: review	Avoidance of verification bias was not associated with test performance	No evidence of bias
Detrano, 1988[79] Type of analysis: statistical Study design: review	Verification bias decreased specificity but did not affect sensitivity	Empirical evidence of bias
Diamond, 1991[104] Type of analysis: narrative Study design: modelling	*Verification bias increased sensitivity and decreased specificity*	*Theoretical evidence of bias*
Diamond, 1992[105] Type of analysis: narrative Study design: modelling	*Verification bias significantly distorts estimates of test performance*	*Theoretical evidence of bias*
Froelicher, 1998[101] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	A study without verification bias showed lower sensitivity and higher specificity than previous studies in which work-up bias was present	Empirical evidence of bias
Lijmer, 1996[106] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	Verification bias can overestimate test performance	Empirical evidence of bias
Lijmer, 1999[33] Type of analysis: statistical Study design: review	Partial verification was not associated with test performance	No evidence of bias
Miller, 1998[107] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	Verification bias overestimated sensitivity and underestimated specificity	Empirical evidence of bias
Mol, 1999[108] Type of analysis: statistical Study design: review	Sensitivity was higher in studies with verification bias than in studies without; specificity was also slightly higher	Empirical evidence of bias
Morise, 1995[85] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	Verification bias increased sensitivity and decreased specificity	Empirical evidence of bias

continued

TABLE 9 *Results of studies that looked at biases associated with verification bias (cont'd)*

Study details	Effect of bias	Category
Morise, 1994[84] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Verification bias increased sensitivity and decreased specificity	Empirical evidence of bias
Panzer, 1987[65] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Verification bias increased sensitivity and decreased specificity	Empirical evidence of bias
Philbrick, 1982[89] Type of analysis: narrative Study design: diagnostic accuracy study, prospective	Verification bias increased sensitivity and decreased specificity	Empirical evidence of bias
Ransohoff, 1982[74] Type of analysis: narrative Study design: review of two studies	Verification bias increased test performance	Empirical evidence of bias
Ransohoff, 1978[66] Type of analysis: narrative Study design: review	Verification bias may overestimate sensitivity	Empirical evidence of bias
Roger, 1997[64] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Verification bias increased sensitivity and decreased specificity	Empirical evidence of bias
Rozanski, 1983[94] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	Decreased specificity is associated with increased verification bias	Empirical evidence of bias
Santana-Boado, 1998[86] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	Verification bias leads to an overestimation of sensitivity and an underestimation of specificity; however, these effects were not significant	No evidence of bias
Zhou, 1994[109] Type of analysis: statistical Study design: modelling with diagnostic accuracy study example	*Theoretical: verification bias leads to biased estimates of test performance* Empirical: positive predictive value is not affected by work-up bias; negative predictive value is effected	Empirical and *theoretical* evidence of bias

TABLE 10 *Results of studies that looked at biases associated with patient inclusion and data collection*

Study details	Effect of bias	Category
Bowler, 1998[102] Type of analysis: statistical Study design: diagnostic accuracy study, retrospective	The study shows considerable scope for verification bias	Empirical evidence of bias
Lijmer, 1999[33] Type of analysis: statistical Study design: review	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	Empirical evidence of bias

TABLE 11 *Results of studies which looked at biases associated with review bias*

Study details	Effect of bias	Category
Detrano, 1989[78] Type of analysis: statistical Study design: review	Review bias was not associated with test performance	No evidence of bias
Detrano, 1988[77] Type of analysis: statistical Study design: review	Blinding reduces the agreement between test results; the effect was significant for sensitivity	Empirical evidence of bias
Detrano, 1988[79] Type of analysis: statistical Study design: review	Blinding reduces the agreement between test results; the effect was significant for sensitivity	Empirical evidence of bias
Lijmer, 1999[33] Type of analysis: statistical Study design: review	Diagnostic review bias overestimated test performance	Empirical evidence of bias
Ransohoff, 1978[66] Type of analysis: statistical Study design: review	Failure to avoid review bias may overestimate sensitivity and specificity	Empirical evidence of bias

Clinical review bias
Nine studies looked at the effects of clinical review bias (*Table 12*). Eight studies found empirical evidence of bias; however, the direction of bias differed between studies. The ninth study found no difference in test performance between those tests interpreted with and without clinical history. Five studies found that the provision of clinical information improved test performance, one study found that more tests were interpreted correctly without clinical history, one study found that test performance was not affected, but that recommendations for further work-up were affected, and the last study found that the effects of clinical history on test performance were variable, but overall test performance was improved.

Observer/instrument variation
Eight studies looked at observer variation; no studies looked at instrument variation (*Table 13*). All studies provided empirical evidence of bias. All eight studies found evidence of interobserver variability and two also found evidence of intraobserver variability; one of these reported that interobserver variability was greater than intraobserver variability. Two studies found that more experienced reviewers, or experts, provided greater sensitivity, while another found that experience was not related to interobserver variability.

Analysis

Precision (sample size, variation by chance)
No studies were identified which provided evidence of the effect of sample size on test performance.

Inappropriate handling of uninterpretable/indeterminate/intermediate test results
Two studies looked at the effects of uninterpretable test results (*Table 14*). 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. The other study found that the treatment of equivocal or non-diagnostic tests was not associated with test performance.

Post hoc choice of threshold value
No studies were identified which provided evidence of the effect of choice of threshold value.

Dropouts
No studies were identified which provided evidence of the effect of dropouts on estimates of test performance.

Summary of results

The sources of bias and/or variation supported by the most empirical evidence were variation by clinical and demographic subgroups, disease prevalence/severity, partial verification bias, clinical review bias and observer/instrument variation (*Table 15*). Other sources of bias and or/variation for which there was some evidence of effect were distorted selection of participants, absence or inappropriate reference standard, differential verification bias and review bias. One

TABLE 12 *Results of studies that looked at biases associated with clinical review bias*

Study details	Effect of bias	Category
Berbaum, 1988[110] Type of analysis: statistical Study design: experimental	Provision of clinical information improved test performance	Empirical evidence of bias
Doubilet, 1981[91] Type of analysis: statistical Study design: experimental	Provision of clinical information improved test performance	Empirical evidence of bias
Eldevick, 1982[111] Type of analysis: statistical Study design: experimental	More tests were interpreted correctly without clinical history than with it	Empirical evidence of bias
Elmore, 1997[112] Type of analysis: statistical Study design: experimental	Knowledge of the clinical history altered the radiologists' level of diagnostic suspicion and overall diagnostic accuracy did improved	Empirical evidence of bias
Froelicher, 1998[101] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	The provision of additional information was found to improve test performance	Empirical evidence of bias
Good 1990[113] Type of analysis: statistical Study design: experimental	There were no statistically significant differences ($p < 0.05$) between cases interpreted with and without clinical history	No evidence of bias
Potchen, 1979[114] Type of analysis: statistical Study design: experimental	Provision of clinical information improved test performance	Empirical evidence of bias
Raab, 2000[115] Type of analysis: statistical Study design: experimental	The presence of clinical history has a variable effect on test performance. Overall, the diagnostic performance of the test improved with the provision of clinical information	Empirical evidence of bias
Schreiber, 1963[116] Type of analysis: statistical Study design: experimental	Provision of clinical information improved test performance	Empirical evidence of bias

study provided evidence that a change in the technology of the index test may bias estimates of test performance and another found that a difference in test protocol may bias estimates of test performance. There was no evidence to support the effects of inappropriate handling of uninterpretable test results on test performance, or disease progression bias. No studies were identified which looked at the effects of incorporation bias, treatment paradox, precision, post hoc choice of threshold value or dropouts.

Discussion

The searches only identified a relatively small number of studies that looked at the effects of bias or variation on estimates of diagnostic test performance. These studies were concentrated in six areas of bias and/or variation: variation by clinical and demographic subgroups (15 studies), disease prevalence/severity (nine studies), absence or inappropriate reference standard (eight studies), verification bias (22 studies), clinical review bias (eight studies) and observer variation (eight studies). Other sources of bias and/or variation commonly believed to affect studies of diagnostic test performance, such as incorporation bias, treatment paradox, post hoc choice of threshold value and dropouts, were not considered in any studies.

The studies included in this review varied in study design and hence quality. The aim of this review was to provide an overview of what types of bias may affect diagnostic test evaluations and so this variation in design is not as much of a problem as it would be in a review that was trying to quantify a specific effect. The aim of the review was not to quantify the extent to which different biases

TABLE 13 *Results of studies that looked at biases associated intraobserver, interobserver or instrument variability of index test*

Study details	Effect of bias	Category
Berbaum, 1989[117] Type of analysis: statistical Study design: experimental	Orthopaedists depend on clinical history more than do radiologists	Empirical evidence of bias
Ciccone, 1992[118] Type of analysis: statistical Study design: experimental	Both inter- and intraobserver variability were found. Interobserver variability was greater than intraobserver variability	Empirical evidence of bias
Cohen, 1987[119] Type of analysis: statistical Study design: experimental	Experts provided higher sensitivity and specificity than non-experts	Empirical evidence of bias
Corley, 1997[120] Type of analysis: statistical Study design: experimental	Some evidence of both inter- and intraobserver variation	Empirical evidence of bias
Cuaron, 1980[121] Type of analysis: statistical Study design: experimental	Very high interobserver variability	Empirical evidence of bias
Elmore, 1994[122] Type of analysis: statistical Study design: experimental	Some evidence of interobserver variability	Empirical evidence of bias
Raab, 1995[123] Type of analysis: narrative Study design: experimental	Some evidence of interobserver variability; this did not appear to be related to experience	Empirical evidence of bias
Ronco, 1996[124] Type of analysis: statistical Study design: experimental	Evidence of interobserver variability, with less experienced examiners showing a lower sensitivity than experienced examiners	Empirical evidence of bias

TABLE 14 *Results of studies that looked at biases associated with uninterpretable, indeterminate and intermediate test results*

Study details	Effect of bias	Category
Detrano, 1989[78] Type of analysis: statistical Study design: review	Treatment of equivocal or non-diagnostic tests was not associated with test performance	No evidence of bias
Philbrick, 1982[89] Type of analysis: statistical Study design: diagnostic accuracy study, prospective	If technically unsatisfactory exercise test results were excluded then 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 were removed from further consideration then 62% would be excluded	No evidence of bias

operate, but to identify which biases may affect test results. All of the included studies fulfil this objective regardless of design or quality. Theoretical evidence and empirical evidence go hand in hand. The advantage of empirical evidence is that it can estimate the amount of bias associated with a particular feature across various situations. It will confirm or disprove the relative importance (ranking) of different features with respect to the potential of bias. If possible, it can pinpoint situations in which a certain bias is more likely to distort measures of diagnostic accuracy. However, sound theoretical principles leading to distorted measures of accuracy will not be disregarded (ignored) simply because an empirical study fails to find evidence for this.

The types of study design used by studies included in the review were classified as reviews, experimental and diagnostic accuracy studies

TABLE 15 *Summary of results*

Category of bias	Source of bias or variation	Evidence of effect of bias (number of studies)		
		Empirical	**Theoretical**	**No evidence**
Spectrum composition	Variation by clinical and demographic subgroups	14	0	1
	Distorted selection of participants	3	0	2
	Disease prevalence/severity	8	1	0
Index test and reference standard				
Selection and execution	Absent or inappropriate reference standard	4	4	0
	Change in technology of index test	1	0	1
	Disease progression bias	0	0	1
	Difference in test protocol	1	0	1
	Partial verification bias	17	3	3
	Differential verification bias	2	0	0
	Incorporation bias	0	0	0
	Treatment paradox	0	0	0
Interpretation	Review bias	4	0	1
	Clinical review bias	7	0	1
	Observer/instrument variation	8	0	0
Analysis	Precision (sample size, variation by chance)	0	0	0
	Inappropriate handling of uninterpretable test results	0	0	2
	Post hoc choice of threshold value	0	0	0
	Dropouts	0	0	0

(retrospective or prospective). Different types of design are more appropriate for different types of bias. For example, to assess the effects of providing clinical information to aid diagnosis or to investigate observer variability an experimental design is the most appropriate, whereas to investigate features such as spectrum composition a diagnostic accuracy study would be more appropriate. The study design of each study included in the review is presented in the tables summarising the results of each study for each source of bias. This provides readers with an indication of the design used.

It is very difficult to draw conclusions as to the direction in which each source of bias will affect the results. The main reason for this is that bias may affect results in different directions in different studies. An attempt was made to provide an indication of the direction of bias in each study in the summary tables; however, this was not always clear from the results of the studies. The direction of each source of bias was also discussed narratively, but it was not possible to draw any overall conclusions regarding the direction in which any particular source of bias affects results.

Similarly, an attempt was made to provide an indication of the amount of bias operating in each study in the summary tables. However, as with the direction of bias, the amount of bias was not always clear from the individual studies and where it was reported it tended to vary considerably between studies. This review was therefore unable to provide an overview of the direction in which and the extent to which each source of bias affects estimates of test performance. What this review does provide is an overview of the different biases that have been evaluated, the number of studies in which they have been evaluated and whether there is any evidence that they may affect estimates of test performance.

Variation by clinical and demographic subgroups, disease prevalence/severity, partial verification bias, clinical review bias and observer/instrument variation were the sources of bias supported by the most empirical evidence of bias. Based on the available evidence, these appear to be the most important biases. There was also some evidence of bias, both empirical and theoretical, for the effects of distorted selection of participants, absent or inappropriate reference standard, differential verification bias and review bias. The evidence for the effects of other sources of bias was insufficient to draw conclusions regarding the effects, if any, of these biases. When interpreting these results it is important to consider the evidence on which they are based. The fact that there is currently no evidence that a particular bias affects estimates of study performance may be because this source of

bias has not been investigated. Studies will be more likely to investigate sources of bias that would theoretically be expected to affect test performance, or that are easy to investigate.

Conclusion

The main objective of this project is to produce a quality assessment tool for the assessment of diagnostic accuracy studies. This review has made an important contribution to this process. It has provided an indication of the evidence available for the effects of each source of bias and/or variation. This will be important information that will help in the selection procedure for items to be included in the quality assessment tool. Based on the results of this review the sources of bias and/or variation for which there is the most evidence of an effect include variation by clinical and demographic subgroups, disease prevalence/severity, partial verification bias, clinical review bias and observer/instrument variation. These need to be confirmed in further real-life empirical studies such as the study by Lijmer and colleagues.[33] Some potential sources of bias have not yet been (sufficiently) evaluated in previous research. Further empirical work is needed to clarify the size and direction of these biases.

Chapter 6

Objective 2: Examine how quality assessment has been handled in systematic reviews

This objective was included in the review to give an overview of how quality assessment has been taken into account in existing systematic reviews of diagnostic tests. Since a quality assessment tool is being developed as part of this project, knowledge about ways to incorporate quality assessments in systematic reviews is relevant, as it will help to determine how the tool will be used in the future. This will have important implications for the structure of the tool.

Methods

Inclusion criteria

To be included in the review systematic reviews had to:

- evaluate the accuracy of a diagnostic or screening test by including studies that compared a test to a reference standard
- conduct any form of quality assessment of the individual studies included in the review.

Studies that used a randomised control design that did not allow the calculation of test performance were excluded, as the quality criteria relevant to these studies are very different to those relevant to standard diagnostic test accuracy evaluations.

Studies were assessed for inclusion by one reviewer and checked by a second reviewer; discrepancies were resolved by consensus.

Literature searches

CRD's DARE database was used to identify existing systematic reviews of diagnostic studies. This database has been compiled from extensive literature searches of a wide range of databases (such as Current Contents, MEDLINE and CINAHL). All abstracts of diagnostic reviews entered onto DARE until April 2001 were eligible for inclusion.

To be included on DARE a systematic review has to meet four of the following six criteria.

- Does the review answer a well-defined question?
- Was a substantial effort made to search for all the relevant literature?
- Are the inclusion/exclusion criteria reported and are they appropriate?
- Is the validity of included studies adequately assessed?
- Is sufficient detail of the individual studies presented?
- Have the primary studies been combined and summarised appropriately?

Data extraction

Data were extracted on:

- author
- year of publication
- diagnostic test evaluated
- reference standard
- target disorder
- study designs included in the review
- whether the review conducted a narrative or statistical synthesis
- whether the quality assessment was designed specifically to look at diagnostic tests
- how the quality assessment was used in the review (*Box 1*)
- the tool used to assess studies.

All identified tools, with the exception of modified tools, were included as part of Chapter 7.

BOX 1 Incorporation of quality assessment into the systematic reviews

The methods in which quality was incorporated into the reviews were classified into the following categories:

- As criteria for inclusion in the review
- As criteria for inclusion in primary analysis
- To conduct sensitivity analyses
- As variables in a regression analysis
- To make recommendations for future research
- As a factor to weight a meta-analysis
- Results presented in a table
- A narrative discussion of quality

TABLE 16 *Quality assessment tools used in the included systematic reviews*

Tool used	Specific tool	No. of studies
Authors' own		41
Modified tool		3
Existing tool	Holleman, 1995[127]	5
	Mulrow, 1989[45]	3
	Irwig, 1994[130]	2
	Sackett, 1991[6]	2
	Cochrane Methods Group, 1996[46]	1
	Kent, 1992[131]	1
	Wilson and Junger criteria for screening programmes[132]	1

The quality assessment tools used were classified as follows:

- existing tool: where an existing tool was used in its published format
- modified tool: where the review authors modified one single existing tool
- author's own: where the review authors either:
  - used and adapted more than one existing tool
  - developed their own tool using standard scale development techniques
  - used a tool but did not reference it or make any statement regarding the origin of the tool.

Data were extracted by one reviewer and checked by a second reviewer; discrepancies were resolved by consensus.

Data synthesis

A narrative synthesis is presented. The proportion of reviews on DARE that assessed study quality was calculated. A general discussion of the criteria used to assess study quality is provided. A more in-depth analysis of the content of the tools is provided in Chapter 7.

Results

The nature of the evidence

In total, 114 systematic reviews of diagnostic or screening tests were identified from DARE. Of these, 58 (51%) conducted some form of quality assessment of the individual studies and were included in this review. None of the studies reported that the validity and reliability of the tools had been estimated. Details of the systematic reviews included in the review are presented in Appendix 3.

Quality assessment tools used in the systematic reviews

The quality assessment tools used in the systematic reviews are summarised in *Table 16*. All tools included in these reviews, including the tools developed by the authors of the review specifically for the review, are discussed further in Chapter 7. The majority of the reviews, 41 (71%), used the author's own tool. None of the studies reported that they had used standard scale development techniques to develop the tool. Fifteen of the reviews based their tools on more than one existing quality assessment tool. The remaining 26 provided no indication of the source of the items included in the tools.

Three reviews used modified tools. One study[125] modified Jaeschke (1994),[35,36] one[126] modified Holleman (1995)[127] and the third[128] modified Mulrow (1989).[45] None of these studies reported how the existing tools had been modified; however, all listed the criteria on which the studies were assessed for methodological quality. In one case,[126] limited modification was made, with the addition of only one item. In another,[125] the original tool was shortened, with the removal of three of the original six items and the addition of one further item. In the final example,[128] quite extensive modifications were made to the original tool: seven of the original items were removed and a further four were added. Modified tools were not included in Chapter 7.

The remaining 13 reviews used existing tools (*Table 16*) and, in addition to using the authors' own criteria, one review used published checklists to assess study quality.[129] The most commonly used published checklist was Holleman (1995),[127] which was used by five reviews, three reviews used Mulrow (1989)[45] and two reviews each used Sackett (1991),[6] and Irwig (1994).[130] Other published checklists, each used in one of the reviews, were the Cochrane Methods Group (1996),[46] Kent (1992)[131] and the Wilson and Junger criteria for screening programmes.[132] The Wilson and Junger criteria are not considered to be criteria for assessing methodological quality of diagnostic tests

TABLE 17 *Quality assessment tools that authors used to produce their own tools*

Published quality assessment tool	Included in Chapter 7?	Review which adapted checklist
Anon, 1981[47]	Yes	Fahey, 1995[134]
Beam, 1991[29]	Yes	Fahey, 1995[134]
Begg, 1988[135]	No: Discussion of different types of bias, not an actual checklist	Lacasse, 1999[136]
Cochrane, 1996[46]	Yes	Chien, 1997[137]
Cooper, 1988[30]	Yes	Fahey, 1995[134]
Deyo, 1994[138]	Yes	van Tulder, 1997[139]
Dunn, 1995[140]	Yes	Chien, 1997[137]
Feinstein, 1985[141]	No: general textbook, not an actual checklist	Fiellin, 2000[142]
Guyatt, 1992[143]	Yes	Chien, 1997[137]
Hoffman, 1991[133]	Yes	van den Hoogen, 1995[144] Owens, 1996[145] van Tulder, 1997[139]
Holleman, 1995[127]	Yes	Badgett, 1997[126]
Irwig, 1994[130]	Yes	Rao, 1995[146] Heffner, 1997[147] Koumans, 1998[148] Nuovo, 1997[149] Fahey, 1995[134] van Tulder, 1997[139]
Jaeschke, 1994[35,36]	Yes	Lacasse, 1999[136] Hobbs, 1997[13] Nuovo, 1997[149] Fiellin, 2000[142]
Kent, 1992[131]	Yes	Adams, 1996[150] van den Hoogen, 1995[144] Owens, 1996[145] van Tulder, 1997[139]
Kent, 1992[151]	Yes	Owens, 1996[145]
Kent, 1994[152]	No: adaptation of Hoffman (1991),[133] Kent (1992)[151] and Kent (1992)[131] and therefore not included separately	Adams, 1996[150]
Meade, 1997[153]	No: quality assessment tool adapted from existing tool, not specifically for diagnostic test evaluations	Fiellin, 2000[142]
Kobberling, 1990[40]	Yes	Fahey, 1995[134]
Mulrow, 1989[45]	Yes	van den Hoogen, 1995[144] Fahey, 1995[134]
Owens, 1996[145,154]	Yes	Heffner, 1997[147] Koumans, 1998[148]
Philbrick, 1980[155]	Yes	Mullins, 2000[156]
Ransohoff, 1978[66]	No: included as objective 1; discusses elements of study quality but does not present a quality assessment tool	Mullins, 2000[156] Fahey, 1995[134]
Reid, 1995[26]	Yes	Hobbs, 1997[13] Nuovo, 1997[149] Fiellin, 2000[142]
Sackett, 1991[6]	Yes	Rao, 1995[146]
van den Hoogen, 1995[144]	Yes	van Tulder, 1997[139]

and so these are not included in Chapter 7. All of the other tools are discussed further in this chapter.

The 13 reviews that adapted more than one existing quality assessment tool to develop their own tool used a total of 25 different tools. The number of tools adapted for each quality assessment tool ranged from two to six. *Table 17* provides details of these tools. The most commonly adapted checklists were Irwig (1994),[130] which was used in six reviews, Jaeschke (1994)[35,36] and Kent (1992)[131], which were used in four reviews, and Hoffman (1991)[133] and Reid (1995),[26] which were used in three reviews. The majority of the published checklists are included as part of objective 3; however, some of the quality assessment tools were not considered to be actual checklists and so these have not been included. Reasons for the exclusion of these tools from Chapter 7 are provided in *Table 17*.

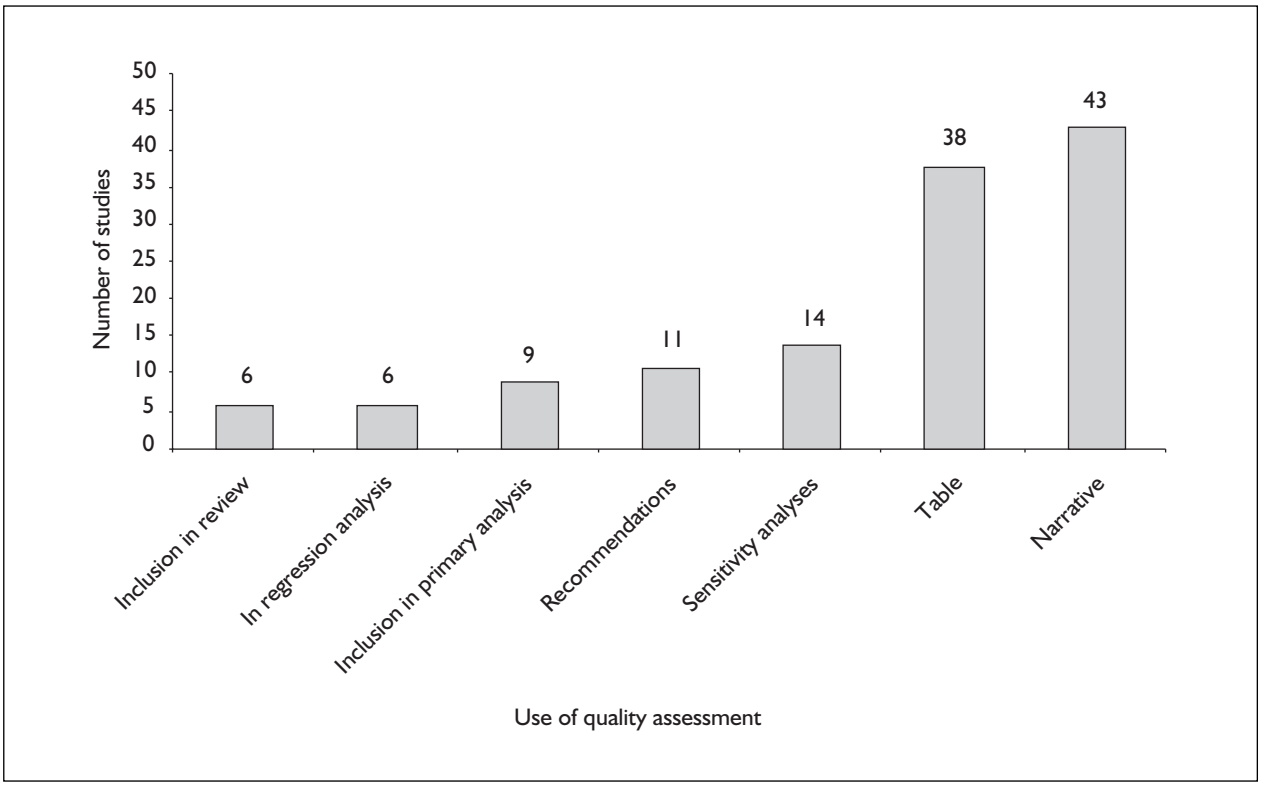

FIGURE 2 *Number of studies that used each method of incorporating quality assessment into the review*

A 'checklist' type quality assessment was used by most reviews (76%). Fourteen reviews (24%) used a 'levels of evidence' approach in which studies were assigned grades or levels according to whether or not they fulfilled certain quality criteria. In three reviews it was not clear which approach was used: one only assessed papers for blinding,[157] a second stated that quality assessment was conducted but provided no details[158] and the third discussed how studies compared to an 'ideal study'.[159] Of the studies that used a checklist nine used the checklist to calculate a quality score for studies. Two of the studies that used a checklist approach also used a level of evidence approach.

Incorporation of quality assessment in the reviews

Table 18 shows how quality assessment was incorporated into the systematic reviews. See *Box 1* for clarification of the different incorporation methods. Note that more than one method of inclusion could be used by any single review. The number of reviews using each method of inclusion is shown in *Figure 2*.

The majority of the reviews identified (43/58) discussed study quality narratively, with 11 using study quality as a basis for recommendations for future research. In 37 reviews the results of the quality assessment were presented in table format.

Eight reviews used part of their quality assessment as inclusion criteria for the review (*Table 19*).[136,139,146,162,177,192,194,196] Seven of these used criteria related to the description of the population,[162,194] method of patient sampling[136,146,177,194] or inclusion of both a diseased and a non-diseased cohort.[139,192] Six reviews used inclusion criteria related to properties of the index and/or reference standard.[136,139,146,162,192,194] One review[177] required that studies met certain design criteria (receipt of reference standard by all patients, independent evaluation of tests and minimum sample size of 50). One study used a 'levels of evidence' approach, stating that all included studies had to be level C or above.[196] The criteria used as inclusion criteria are summarised in *Table 19*.

Several other studies specified quality-related variables as inclusion criteria; however, this did not form any part of the quality assessment. The majority of these (n = 16) only included studies that compared the test of interest to a specific reference standard.[128,134,157,159–161,163,168,169,173,175,181,182,185,187,193] Three others only included cohort studies[172] or prospective studies,[188] or required consecutive enrolment of patients.[171] Fifteen reviews only included studies that provided sufficient information for the construction of 2 × 2 tables.[126,128,134,145,154,165,166,172,173,175,181,185,186,190,193]

TABLE 18 *How quality assessment was incorporated into the systematic reviews*

Study details	Inclusion in review	Inclusion in primary analysis	Sensitivity analyses	In regression analysis	Recommendations	Table	Narrative
Adams, 1996[150]	✗	✗	✗	✗	✗	✓	✓
Anand, 1998[160]	✗	✗	✗	✗	✗	✗	✗
Attia, 1999[161]	✗	✗	✗	✗	✗	✓	✓
Bachmann, 1998[162]	✓	✗	✗	✗	✗	✗	✗
Badgett, 1996[163]	✗	✗	✗	✗	✗	✗	✓
Badgett, 1997[126]	✗	✗	✗	✗	✗	✓	✗
Barlow, 1998[164]	✗	✗	✗	✗	✗	✓	✓
Bastian, 1997[165]	✗	✗	✗	✗	✗	✓	✗
Bastian, 1998[166]	✗	✗	✗	✗	✗	✓	✓
Becker, 1996[167]	✗	✓	✗	✗	✓	✓	✓
Bell, 1998[168]	✗	✗	✗	✗	✗	✗	✓
Bonis, 1997[169]	✗	✗	✓	✗	✗	✗	✓
Bradley, 1998[170]	✗	✗	✗	✗	✗	✓	✗
Buntinx, 1997[171]	✗	✗	✗	✗	✗	✗	✗
Chien, 1997[137]	✗	✗	✓	✗	✗	✓	✓
Conde-Agudelo, 1998[172]	✗	✗	✗	✗	✗	✗	✓
Da Silva, 1995[173]	✗	✗	✗	✗	✗	✗	✓
De Bernardinis, 1999[174]	✗	✗	✓	✗	✗	✗	✓
de Vries, 1996[175]	✗	✗	✗	✓	✗	✓	✓
Devous, 1998[157]	✗	✗	✗	✓	✗	✓	✓
Fahey, 1995[134]	✗	✗	✓	✓	✓	✓	✓
Fiellin, 2000[142]	✗	✗	✓	✗	✓	✓	✓
Hallan, 1997[158]	✗	✗	✗	✗	✗	✗	✗
Heffner, 1995[129]	✗	✗	✗	✗	✗	✗	✓
Heffner, 1997[147]	✗	✗	✗	✗	✗	✗	✓
Hobbs, 1997[13]	✗	✗	✓	✗	✓	✓	✓
Hrung, 1999[176]	✗	✗	✗	✗	✓	✓	✓
Huicho, 1996[128]	✗	✗	✗	✗	✗	✓	✓
Kearon, 1998[177]	✓	✗	✗	✗	✓	✗	✗
Koelemay, 1996[178]	✗	✗	✗	✗	✗	✓	✓
Koumans, 1998[148]	✗	✗	✗	✗	✗	✓	✓
Lacasse, 1999[136]	✓	✗	✗	✗	✗	✓	✓
Lederle, 1999[179]	✗	✗	✗	✗	✗	✗	✓
Liedberg, 1996[180]	✗	✗	✓	✗	✗	✗	✓
Littenberg, 1995[181]	✗	✓	✗	✗	✓	✗	✓
Loy, 1996[125]	✗	✗	✗	✓	✗	✓	✓
Mayer, 1997[182]	✗	✗	✗	✗	✗	✓	✓
McGee, 1999[183]	✗	✗	✗	✗	✗	✓	✗
Metlay, 1997[184]	✗	✓	✗	✗	✗	✗	✗
Mol, 1997[185]	✗	✗	✗	✓	✗	✗	✗
Mol, 1998[186]	✗	✓	✗	✓	✗	✓	✓
Mol, 1998[187]	✗	✓	✗	✓	✗	✓	✓
Mullins, 2000[156]	✗	✗	✗	✗	✗	✓	✓
Nuovo, 1997[149]	✗	✗	✗	✗	✓	✓	✓

continued

TABLE 18 *How quality assessment was incorporated into the systematic reviews (cont'd)*

Study details	Inclusion in review	Inclusion in primary analysis	Sensitivity analyses	In regression analysis	Recommendations	Table	Narrative
Owens, 1996[145,154]	✗	✗	✓	✗	✓	✓	✓
Pearl, 1996[188]	✗	✗	✗	✗	✓	✓	✓
Pollitt, 1997[189]	✗	✗	✗	✗	✗	✓	✗
Rao, 1995[146]	✓	✗	✓	✗	✗	✗	✓
Rappeport, 1996[159]	✗	✗	✓	✗	✗	✓	✗
Reed, 1996[190]	✗	✗	✗	✗	✗	✓	✓
Selker, 1997[191]	✗	✗	✗	✗	✗	✓	✗
Spencer-Green, 1997[192]	✓	✓	✓	✗	✗	✓[a]	✓
Swart, 1995[193]	✗	✗	✓	✗	✗	✓	✓
Tugwell, 1997[194]	✓	✗	✗	✗	✗	✗	✗
van den Hoogen, 1995[144]	✗	✗	✗	✗	✓	✓	✓
van Tulder, 1997[139]	✓	✓	✗	✗	✗	✓	✓
Wells, 1995[195]	✗	✗	✓	✗	✗	✓	✓
Whited, 1998[196]	✓	✗	✗	✗	✗	✗	✓
Total	8	7	13	7	11	37	43

[a] This study presented results in a histogram rather than a table.

TABLE 19 *Quality assessment criteria also used as inclusion criteria for the reviews*

Quality aspect	Inclusion criterion
Description of study population and setting	Defined population[162] Inclusion of both diseased and non-diseased cohort[139,192] Provision of clinical details of patients[194] Specification of patient selection method[146,194] Consecutive patients[136,177]
Test properties	Definition of test performance[162,192,194] Listed reference standard[162] Sufficient details of the reference standard[194] Appropriate reference standard[136,139,146,192]
Study design	Independent evaluation of tests[177] Receipt of reference standard by all patients[177] Minimum sample size of 50[177]
Other	'Levels of evidence' approach[196] Sufficient data for 2 × 2 table[136,146,194]

Seven reviews used study quality as criteria for inclusion in the primary analysis.[139,167,178,181,184,186,187] Four of these reviews graded studies into levels according to methodological quality, including only those which achieved the higher levels of evidence in the primary analysis.[139,167,178,184] One review only included studies that compared the test of interest to a reference standard and provided sufficient information for the construction of a 2 × 2 table.[181] Two reviews only included diagnostic cohort studies in their primary analyses.[186,187]

Thirteen reviews used study quality to conduct sensitivity analyses. Seven identified 'higher quality' studies and either discussed their results separately[13,159,180] or compared their results with those of the lower quality studies[146,174,195] or all studies combined.[145] The other six reviews looked at study results stratified according to specific aspects of methodological quality, including control group used,[192] presence of verification bias[134,142,169] or review bias,[134,142] and various other criteria.[134,137,193]

Seven reviews investigated the effects of various quality-related variables on test performance by including these as variables in a regression analysis.[125,134,157,175,185–187] The most commonly investigated variables were sampling method (four reviews),[125,185–187] reference standard used (four reviews)[125,134,175,185] and blinding (five reviews).[125,134,157,175,185] Other investigated variables were proportion with disease,[125] patient characteristics,[125] study design,[186,187] verification bias,[134] sample size,[125] data collection[186,187] and missing data.[175]

None of the systematic reviews used the quality assessment to weight the meta-analysis. Two reviews stated that a quality assessment had been performed, but did not present any results for this.

Discussion

The main strength of this review was the availability of a large number of systematic reviews for assessment provided via DARE. This database is fed monthly, and in some cases weekly, by extensive literature searches of a wide range of databases (such as Current Contents, MEDLINE, EMBASE and CINAHL). Potential systematic reviews have to meet four out of six quality criteria for systematic reviews before being included on the database. As a consequence, the systematic reviews included here are a sample of the better quality reviews available, and it is likely that if all systematic reviews of diagnostic tests were evaluated then a worse picture of the use of validity assessment would emerge.

Of the reviews identified from DARE only a relatively small proportion (51%) conducted any form of quality assessment. Of those that did, the vast majority (72%) developed their own quality assessment tool rather than using a previously published tool. None of these used standard scale development techniques, although around half stated that they used existing tools as the basis for their own one.

The majority of reviews limit the incorporation of quality in the review to a narrative discussion or the presentation of results in a table.

Three of the reviews did not report or use the results of the quality assessment in any way and almost one-third of the reviews did no more than summarise the results in a table or present a narrative discussion. Only 69% of the reviews used the results of the quality assessment in the study synthesis. Less than one-quarter of reviews that conducted a quality assessment used quality-related variables to conduct sensitivity analyses or to make recommendations for future research. Even fewer reviews went further, using quality-related factors as a basis for inclusion of studies in the review or in primary analyses, or as variables in a regression analysis.

From the point of view of developing a quality assessment tool for diagnostic test evaluations, this review has provided an indication of how quality is currently incorporated into systematic reviews. This review has suggested that there is some need for the tool to be used for conducting sensitivity analyses, to make recommendations for future research, as criteria for including studies in a review or in primary analyses, and to be used in regression analyses. None of the studies used quality-related variables to weight meta-analyses and so this is unlikely to be a requirement of future tools.

When developing the quality assessment tool it is important that results can be discussed narratively, reported in a table or to make recommendations for future research; this will be possible for almost any tool and will therefore not directly impact on the development of the tool. The requirement that the tool needs to be used as criteria for inclusion of studies in the review or in primary analyses has a number of implications for the development of the tool. The tool needs to be able to distinguish between high- and low-quality studies so that inclusion in the review or primary analysis can be restricted to those of higher quality. There are three ways in which this can be done. It may be important to highlight quality-related factors that have the greatest potential to lead to biased results. These may then be used as criteria for inclusion in the review, while other criteria, for which there is some evidence that they may bias results, could be assessed in those that meet the 'main criteria'. Another option would be to produce a tool that can be adapted into a 'levels of evidence' approach, so that the criteria within the tool can be used to stratify studies into different quality levels so that only those that reach the highest quality levels are included in the review or the primary analyses. Alternatively, a quality scoring system can be used whereby the review or primary analyses can be limited to those achieving a certain quality score.

The tool also needs to be able to be used to conduct sensitivity and regression analyses. All of

the above approaches could also be adapted for these requirements. Studies that meet certain quality criteria could be compared with those that do not meet these criteria in a sensitivity analysis, or studies of one level could be compared with those of other levels, and studies above a certain quality score could be compared with those below this score. Similarly, in a regression analysis, the level of evidence or the quality score could be included as explanatory variables. Both sensitivity analyses and regression analyses can go a step further and look at the effects of individual quality-related factors. Thus, quality-related factors that have the highest possibility of bias could be investigated to see whether these show any association with test performance.

In summary, the quality assessment tool needs to have the potential to be discussed narratively, reported in a tabular summary, used in formulating recommendations for future research, used to conduct sensitivity or regression analyses and used as criteria for inclusion in the review or a primary analysis. The resulting implication for the development of the tool is that some distinction needs to be made between high- and low-quality studies. This may be done by highlighting criteria that are more likely to lead to bias, developing a levels of evidence approach and using a quality score. There are advantages and disadvantages associated with all of these methods. These are discussed further in Chapter 8, as part of the development of the quality assessment tool.

Although the main purpose of conducting this review was to provide an overview of how quality assessment has been incorporated into existing reviews of diagnostic tests to help with the development of the quality assessment tool, it has also provided important information on the limitations of reviews of diagnostic tests. This review has revealed that the conduct of systematic reviews of diagnostic test studies with respect to quality assessment is similar to that found for meta-analyses of RCTs[197] and systematic reviews that include non-randomised studies.[198] The authors of the review of randomised studies found that RCT quality is not assessed in almost half of meta-analyses (48%). In those meta-analyses that do assess quality, most use non-validated tools and the results are infrequently incorporated into the analyses (25% of reviews).[197] For non-randomised studies of therapeutic interventions, only 30% of systematic reviews assessed quality, 42% using their own quality assessment tool. Only 37% attempted to incorporate the results of the quality assessment in a quantitative way and 12% did not incorporate the results of the quality assessment into the review synthesis at all.[198]

Conclusion

Reviewers who conduct systematic reviews of diagnostic test evaluation studies should be aware of the fundamental need to assess the quality of the included studies and to examine study quality as a potential source of heterogeneity in the results of their reviews. Furthermore, there is a clear need for a new quality assessment tool for diagnostic test studies to be developed using standard scale development techniques. Such a tool could then be validated by use in future systematic reviews of diagnostic test studies.

Chapter 7

Objective 3: Examine existing methods or assessment tools that have been used to assess the quality of diagnostic research, and any evidence on which they are based

This chapter will provide an indication of the criteria included in existing tools to assess the quality of diagnostic test evaluations. This will help in selecting items for inclusion in the quality assessment tool being developed as part of this project.

Methods

Inclusion criteria

All published checklists used either to assess the quality of diagnostic test studies (e.g. those included in systematic reviews), as guides for reporting such studies or as guides for interpreting reports of diagnostic test studies were included. Inclusion was assessed by one reviewer and checked by a second; discrepancies were resolved by consensus or discussion with a third reviewer where necessary. Each quality assessment tool was only included in the review once. Where duplicate reports of the same tool were found, the publication in which the tool was first used was included.

Literature searches

The search strategies developed for objective 1 were also used to search for assessment tools that have been used to assess the quality of diagnostic research. Quality assessment tools used in systematic reviews were identified as part of objective 2.

Data extraction

Data were extracted on:

- author
- year of publication
- aim of scale: to assess study quality, guides for reporting or interpreting diagnostic tests
- type of scale: checklist, levels of evidence or quality score
- source of tool: original tool, modified tool, authors' own developed for a review
- items addressed by scale: a list of all possible items covered by the scales was produced and each scale was assessed to see which of these items it addressed
- how items were chosen for inclusion on the scale
- topic area for which scale was developed
- time taken to complete scale
- level of inter-rater reliability.

Data were extracted by one reviewer and checked by a second reviewer; discrepancies were resolved by consensus or consultation with a third reviewer.

Data synthesis

Methods used for the development of the scales, the topic areas for which the scales were developed, the level of inter-rater reliability and the time taken to complete each of the scales are discussed. The items covered by the scales were classified as follows:

Spectrum composition
- Spectrum composition
- Inclusion criteria
- Population recruitment
- Disease prevalence/severity

Index test and reference standard

Selection and execution
- Absent or inappropriate reference standard
- Change in technology of index test
- Disease progression bias
- Test execution
- Reference execution
- Verification bias
- Incorporation bias
- Normal defined
- Treatment paradox

Interpretation
- Review bias
- Clinical review bias
- Observer/instrument variation

Data presentation
- Appropriate results
- Precision (sample size, variation by chance)
- Inappropriate handling of uninterpretable/ indeterminate/intermediate test results
- Post hoc choice of threshold value
- Dropouts
- Subgroups
- Data table
- Utility of test

Research planning
- Sample size
- Objectives
- Protocol

Detailed definitions of each of these items are provided in Chapter 4. A table was produced to describe which of the criteria each individual quality assessment tool covered. The proportion of tools assessing each item was calculated. Each item was then classified from I to IV according to the proportion of scales in which each particular item was included:

Classification	Proportion of scales in which the item was included
I	75–100%
II	50–74%
III	25–49%
IV	0–24%

Results

The nature of the evidence

In total, 91 quality assessment tools were eligible for inclusion in the review: 40 quality assessment tools were identified as part of objective 2 and details of these tools are presented in Appendix 3. Literature searches identified a further 51 quality assessment tools, details of which are presented in Appendix 4. Overall, 58 of the 91 tools were authors' own tools, developed for use in systematic or methodological reviews.[13,26,27,29,30,75,127,129,131,133,134,136,137,139,142–151,155–159,161,163,167,170–178,181,184–187,189,190,193,195,199–206] A further five tools were original tools developed to assess the quality of studies included in systematic reviews,[46,130,207–209] and four additional tools were developed to assess the quality of studies of diagnostic accuracy.[45,210–212] The authors of these tools did not specifically state that they were developed to be used in systematic reviews, although they could be used for this purpose. Of the remaining 24 tools, 21 were guides for the interpretation (12 tools),[6,35,40,47,48,140,183,213–217] conduct (six tools),[4,138,218–221] or reporting (three tools)[222–224] of studies of diagnostic accuracy, and three tools[1,67,225] were lists of biases that may affect studies of diagnostic accuracy.

Tool development and methodological details

The majority (66%) of the tools did not report how quality items were selected for inclusion. Twenty-five tools stated that they were adapted from previous tools, but did not give any further details on how this occurred.[13,29,48,126,130,133,134,136,137,139,142,144–150,155,156,199,203,206,212,216] Four tools stated that they were developed from established literature on diagnostic test evaluations, but the authors of these tools also failed to provide any further information on how items were selected for inclusion.[27,30,75,208] Only two quality assessment tools provided detailed information on how items were selected for inclusion in the tool. One was developed by examining published reports on the shortcomings of studies of diagnostic accuracy, preparing an initial draft checklist and presenting this at a meeting of editors. Comments from these editors were then incorporated into the tool and a revised version was published.[222] The other was developed through a process involving 14 panel members, all of whom had practical experience in using diagnostic tests and nine of whom had training in clinical epidemiology.[45] A series of five steps was used to identify and weight questions for inclusion in the tool, and during a final process the items to be included and the weights to be given to them were finalised. Further details are presented in Appendix 4.

None of the tools reported how long it would take to complete the scale. Only eight reports of the tools contained information on the level of inter-rater reliability of the tool.[27,45,133,143,199,201,203,208] Kappa values for inter-rater reliability ranged from 0.26 to 0.92.

Most tools used a 'checklist' type quality assessment (67%), eight used a level of evidence approach and 12 were used to produce an overall quality score. One review only assessed papers for blinding[157] and a second discussed how studies compared to an 'ideal study'.[159]

All of the tools developed within the context of systematic reviews or methodological reviews of diagnostic test studies were developed for specific topic areas. However, only six of these tools included topic-specific items[134,137,148,163,170,193] (see *Table 20*). These tools were all developed for different topic areas and included a variety of

TABLE 20 *Topic-specific criteria included in the quality assessment tools*

Study details	Topic area	Topic-specific criteria
Badgett, 1996[163]	Radiography	Were the radiographs posteroanterior films?
		Was the radiograph interpreted by an experienced radiologist or a cardiologist?
Bradley, 1998[170]	Alcohol screening questionnaires	Items were used in more than one questionnaire, sometimes with changes in time-frame or wording, potentially limiting generalisability
		Multiple alcohol screening questionnaires were administered at one time, potentially leading to consistency response bias or other context response biases
		Screening questionnaires and comparison standards were administered by the same interviewer or at one sitting, potentially biasing questionnaire and interview responses towards higher agreement
		Criterion standards were not interview administered, potentially affecting their validities
Chien, 1997[137]	Test for preterm delivery	Assessment of gestational age: ideal: based on date of last menstrual period confirmed with ultrasound scan before 20 weeks of gestation; second best: in absence of menstrual date early pregnancy scans were performed to confirm gestational age; unclear: did not provide any information
Fahey, 1995[134]	Pap test for cervical precancer	Clinical use: follow-up test if prompted by findings in previous Pap test, otherwise characterised as screening
		Technique described: if technique used to collect cervical cells was reported
Koumans, 1998[148]	Polymerase chain reaction for gonorrhoea	If gender of participants was not described or there were fewer than five culture-positive participants performance results were neither abstracted nor combined unless specimens were taken from pharynx or rectum
Swart, 1995[193]	Hysterosalpingography for tubal pathology	Contrast use (oil vs water) Presence of spasmolyticum (yes or no)

topic-specific criteria, making it difficult to draw any conclusions regarding particular methodological criteria appropriate for inclusion in a topic-specific section of a tool.

The majority of the remaining tools were developed for a general setting. The remainder were developed for specific-topic areas. Eight tools were developed for imaging studies.[4,29,131,155,205,208,210,220] Other topic areas each covered by one tool were dementia,[219] veterinary medicine[212] and pleural cavity.[213] None of these tools contained elements specific to the topic area under study: all of these tools could be used as a quality assessment tool for diagnostic studies in any topic area.

Items included in the quality assessment tools

The items included in the quality assessment tools are shown in *Table 21*. Tools highlighted in grey were identified as part of objective 2; all other tools were identified from literature searches for this section. The proportion of studies covering each quality criterion, grouped according to quality category, is illustrated in *Figure 3*.

Spectrum composition

Eighty-two of the 91 (90%) tools included at least one criterion related to the description of the study population or setting. Spectrum composition was the most commonly included criterion and was included in 64% of tools. Population recruitment was included in 58% of tools, while inclusion criteria and disease prevalence were included in 12% and 10% of tools, respectively.

Index test and reference standard

Selection and execution

Eighty-five studies (92%) included at least one criterion related to the selection and execution of the index test and reference standard. The most frequently included criterion was the use of an appropriate reference standard (64% of tools). Verification bias was also included in a high proportion of tools (63%). The provision of a description of the execution of the index test and the definition of the cut-off point for a normal or an abnormal test were included in 40% and 46% of tools, respectively. The provision of an appropriate description of reference standard execution was included in 11% of tools. Other

TABLE 21 *Criteria covered by the quality assessment tools*

Study details	Spectrum composition				Index test and reference standard												Data presentation							Research planning			
					Selection and execution								Inter-pretation															
	Spectrum composition	Inclusion criteria	Population recruitment	Disease prevention/severity	Reference standard	Change in technology	Disease progression bias	Test execution	Reference execution	Verification bias	Incorporation bias	Normal defined	Treatment paradox	Review bias	Clinical review bias	Observer/instrumentation variability	Appropriate results	Precision of results	Indeterminate results	Post hoc choice of threshold	Dropouts	Subgroups	Data table	Utility of test	Sample size	Objectives	Protocol	Total
Adams, 1996[150]	✓				✓									✓											✓			4
Anon, 1981[47]	✓		✓		✓			✓				✓		✓		✓	✓						✓	✓				10
Arrive, 2000[208]	✓	✓	✓		✓			✓		✓	✓	✓		✓		✓	✓		✓							✓		13
Attia, 1999[161]					✓					✓																		2
Badgett, 1996[163]	✓						✓					✓		✓														4
Becker, 1996[167]	✓		✓	✓	✓			✓	✓	✓				✓		✓	✓											10
Beam, 1991[29]	✓	✓	✓		✓									✓	✓	✓	✓	✓							✓		✓	11
Becker, 1989[206]	✓		✓					✓		✓		✓		✓		✓												7
Black, 1990[210]	✓		✓		✓			✓		✓		✓		✓			✓						✓					9
Bradley, 1998[170]	✓		✓		✓					✓							✓								✓			6
Bruns, 2000[222]	✓	✓	✓					✓	✓	✓		✓	✓	✓		✓	✓	✓	✓				✓		✓			15
Buntinx, 1997[171]	✓		✓																						✓			3
Chien, 1997[137]			✓									✓		✓							✓							4
Cochrane, 1996[46]	✓		✓	✓	✓			✓		✓		✓	✓	✓	✓	✓			✓						✓			13
Conde-Agudelo, 1998[172]			✓					✓		✓		✓		✓			✓			✓								7
Cooper, 1988[30]					✓									✓		✓	✓						✓				✓	6
Da Silva, 1995[173]	✓		✓		✓					✓																		4
De Bernardinis, 1999[174]			✓		✓															✓								3
Deeks, 1999[216]	✓				✓					✓	✓			✓				✓										6
Deeks, 2001[1]	✓		✓		✓			✓	✓	✓		✓	✓	✓					✓									10
de Vries, 1996[175]					✓					✓				✓			✓											4
Devous, 1998[157]														✓														1
Deeks, 2001[225]	✓		✓		✓			✓	✓	✓	✓		✓	✓					✓									10
Deyo, 1994[138]	✓		✓		✓			✓			✓			✓	✓	✓	✓							✓				10
Dunn, 1995[140]	✓		✓		✓			✓						✓		✓	✓	✓						✓	✓			10
Fahey, 1995[134]			✓							✓		✓		✓			✓	✓		✓								7
Fiellin, 2000[142]	✓									✓				✓								✓						4
Freedman, 1987[4]	✓				✓			✓		✓	✓	✓		✓	✓	✓	✓		✓						✓			12
Gifford, 1999[219]	✓									✓				✓		✓	✓											5
Greenhalgh, 1997[48]	✓				✓					✓		✓		✓		✓		✓						✓				8
Greiner, 2000[212]	✓	✓	✓		✓			✓	✓			✓		✓			✓	✓	✓				✓	✓	✓	✓	✓	16
Guyatt, 1992[143]			✓					✓						✓														3
Hallan, 1997[158]	✓			✓	✓			✓		✓				✓						✓					✓			8
Haynes, 1995[224]	✓				✓									✓														3
Heffner, 1998[199]	✓		✓		✓					✓		✓		✓		✓	✓	✓	✓			✓		✓				12
Heffner, 1998[213]	✓		✓		✓	✓			✓	✓		✓		✓		✓	✓	✓	✓			✓						13
Heffner, 1995[129]					✓					✓				✓														3
Heffner, 1997[147]	✓				✓			✓		✓				✓										✓				6
Hobbs, 1997[13]	✓				✓			✓		✓				✓			✓							✓				7

continued

TABLE 21 *Criteria covered by the quality assessment tools (cont'd)*

Study details	Spectrum composition				Index test and reference standard											Data presentation							Research planning			
					Selection and execution								Inter-pretation														
	Spectrum composition	Inclusion criteria	Population recruitment	Disease prevention/severity	Reference standard	Change in technology	Disease progression bias	Test execution	Reference execution	Verification bias	Incorporation bias	Normal defined	Treatment paradox	Review bias	Clinical review bias	Observer/instrumentation variability	Appropriate results	Precision of results	Indeterminate results	Post hoc choice of threshold	Dropouts	Subgroups	Data table	Utility of test	Sample size	Objectives	Protocol	Total
Hoffman, 1991[133]	✓		✓		✓					✓				✓										✓			6
Holleman, 1995[127]			✓		✓					✓				✓													4
Hrung, 1999[176]				✓						✓				✓			✓							✓			5
Irwig, 1994[130]					✓					✓				✓													3
Jaeschke, 1994[35]	✓				✓			✓		✓				✓	✓		✓							✓				8
Jensen, 1999[211]	✓				✓			✓		✓		✓		✓				✓	✓			✓		✓	✓	✓	✓	13
Kelly, forthcoming[205]	✓	✓	✓				✓			✓			✓	✓		✓			✓									9
Kearon, 1998[177]			✓							✓				✓										✓			4
Kent, 1992[131]	✓				✓					✓				✓					✓					✓			6
Kent, 1992[151]	✓		✓		✓			✓		✓	✓			✓										✓			8
Khan, 2001[209]			✓		✓					✓		✓		✓													5
Kobberling, 1990[40]	✓	✓		✓						✓				✓			✓						✓			✓	8
Koelemay, 1996[178]	✓		✓		✓			✓				✓		✓													6
Koumans, 1998[148]	✓		✓					✓	✓	✓				✓	✓									✓			8
Kraemer, 1992[218]	✓		✓		✓			✓				✓		✓			✓		✓		✓			✓			11
Lacasse, 1999[136]			✓							✓				✓													3
Lang, 1997[223]	✓				✓			✓		✓		✓		✓		✓	✓	✓	✓			✓		✓	✓			13
Lensing, 1993[200]	✓		✓					✓				✓		✓		✓												6
Liddle, 1996[207]	✓				✓					✓				✓							✓							5
Littenberg, 1995[181]			✓		✓									✓										✓			4
Mant, 1995[217]					✓																		✓			✓	3
McGee, 1999[183]					✓					✓				✓										✓			4
Metlay, 1997[184]			✓		✓									✓										✓			4
Mol, 1997[185]			✓		✓									✓													3
Mol, 1998[186]			✓																								1
Mol, 1998[187]			✓		✓					✓				✓													4
Mower, 1999[67]	✓			✓	✓	✓				✓		✓		✓	✓	✓			✓			✓					11
Mullins, 2000[156]	✓		✓	✓				✓		✓		✓		✓		✓											8
Mulrow, 1989[45]	✓	✓			✓			✓	✓			✓		✓		✓	✓		✓					✓			11
Nuovo, 1997[149]	✓		✓					✓		✓		✓				✓								✓			7
Owens, 1996[145,154]	✓		✓		✓			✓		✓				✓	✓										✓		8
Panzer, 1986[75]	✓		✓		✓				✓	✓				✓		✓	✓										8
Philbrick, 1980[155]	✓									✓				✓													3
Pollitt, 1997[189]			✓																								1
Radack, 1993[201]	✓		✓		✓					✓				✓		✓	✓										7
Rao, 1995[146]			✓											✓			✓										3
Rappeport, 1996[159]			✓							✓				✓													3
Reed, 1996[190]		✓														✓										✓	3
Reid, 1995[26]	✓									✓				✓		✓		✓	✓			✓					7

continued

TABLE 21 *Criteria covered by the quality assessment tools (cont'd)*

Study details	Spectrum composition				Index test and reference standard												Data presentation								Research planning			
					Selection and execution									Inter-pretation															
	Spectrum composition	Inclusion criteria	Population recruitment	Disease prevention/severity	Reference standard	Change in technology	Disease progression bias	Test execution	Reference execution	Verification bias	Incorporation bias	Normal defined	Treatment paradox	Review bias	Clinical review bias	Observer/instrumentation variability	Appropriate results	Precision of results	Indeterminate results	Post hoc choice of threshold	Dropouts	Subgroups	Data table	Utility of test	Sample size	Objectives	Protocol	Total
Riegelman, 1996[214]	✓				✓			✓				✓				✓	✓					✓						7
Rothwell, 2000[202]	✓		✓					✓				✓		✓	✓	✓			✓						✓			9
Sackett, 1991[6]	✓		✓		✓			✓				✓		✓		✓								✓				8
Sackett, 2000[215]	✓				✓					✓				✓		✓												5
Sheps, 1984[27]				✓	✓							✓		✓			✓						✓					6
Sox, 1989[221]	✓				✓					✓		✓		✓		✓						✓						7
Swart, 1995[193]			✓	✓			✓					✓		✓											✓			6
Thornbury, 1991[220]	✓									✓	✓			✓	✓										✓			6
van den Hoogen, 1995[144]	✓	✓	✓					✓	✓	✓				✓									✓		✓			9
van der Wurff, 2000[203]	✓	✓					✓	✓				✓		✓			✓				✓				✓			9
van Tulder, 1997[139]		✓	✓		✓			✓		✓		✓		✓		✓					✓		✓					10
Windeler, 1988[204]					✓			✓			✓			✓			✓						✓	✓	✓	✓		9
Wells, 1995[195]			✓		✓					✓		✓		✓														5
Total	59	11	53	9	59	2	4	37	10	58	8	35	5	80	10	32	32	12	18	3	6	9	11	15	30	6	5	

criteria, disease progression bias, incorporation bias, treatment paradox and a change in the technology of the index test, were included in less than 10% of tools.

Interpretation

Eighty-two of the 91 tools (95%) included at least one criterion related to interpretation. The most frequently included criterion was review bias (87% of tools). In most cases this related to test review bias or diagnostic review bias, that is, where knowledge of one test result influences the interpretation of the other. Clinical review bias, where test interpretation is influenced by clinical information, was mentioned by only 11% of tools. Observer/instrument variability was covered by 35% of tools.

Data presentation

The quality category data presentation was included in 61% of tools. The most commonly included criterion was the presentation of appropriate results (35% of tools). The handling of uninterpretable/indeterminate/intermediate test results was featured in 18% of tools, test utility in 16%, precision of results in 13% and the presentation of a 2 × 2 results table in 12% of tools. Other items, post hoc choice of threshold, dropouts and subgroup analyses were included in less than 10% of tools.

Research planning

Research planning was the quality category included in the least number of scales. Only 39% of tools included at least one item related to research planning. The most commonly included item in this category was sample size (33% of tools). The provision of a clear description of the study's objectives and the availability of a study protocol were each included in less than 10% of tools.

Number of categories covered by each quality assessment tool

The tools used in the reviews varied in complexity and the number of quality categories covered. The

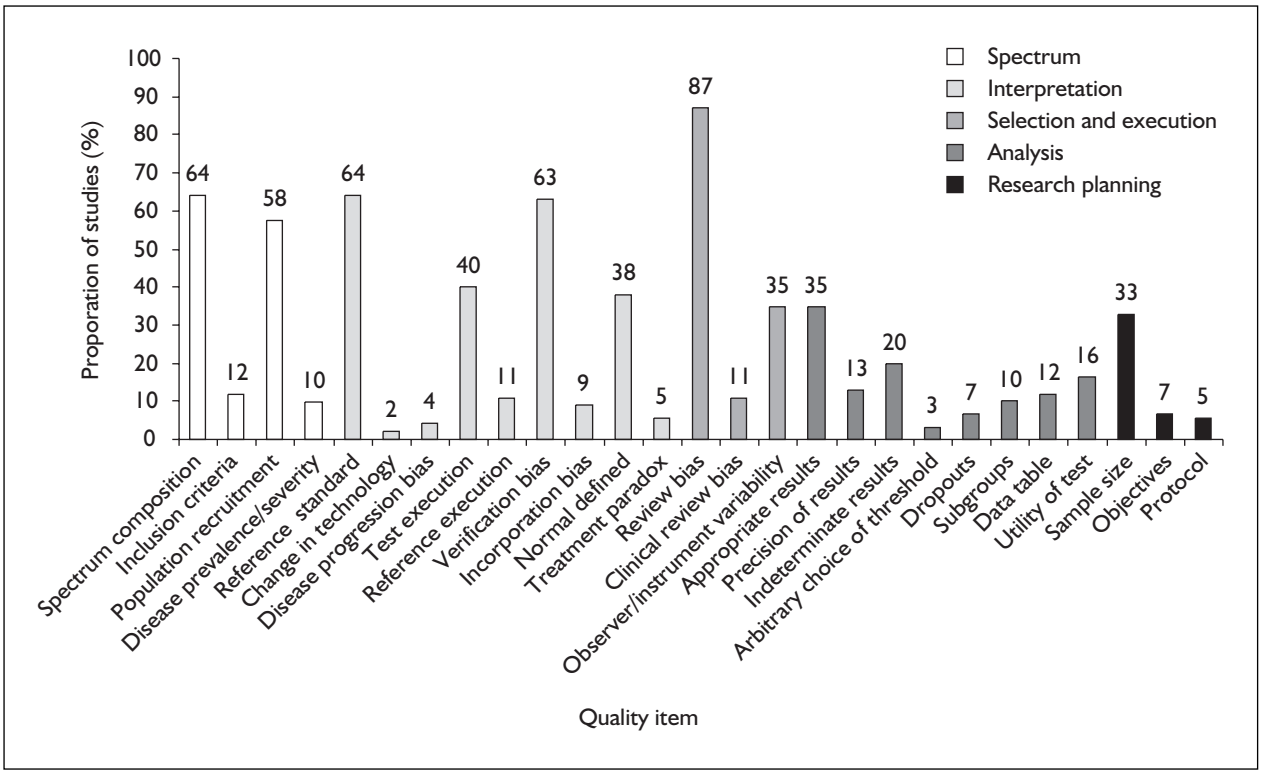

FIGURE 3 *Proportion of quality assessment tools covering each quality criterion*

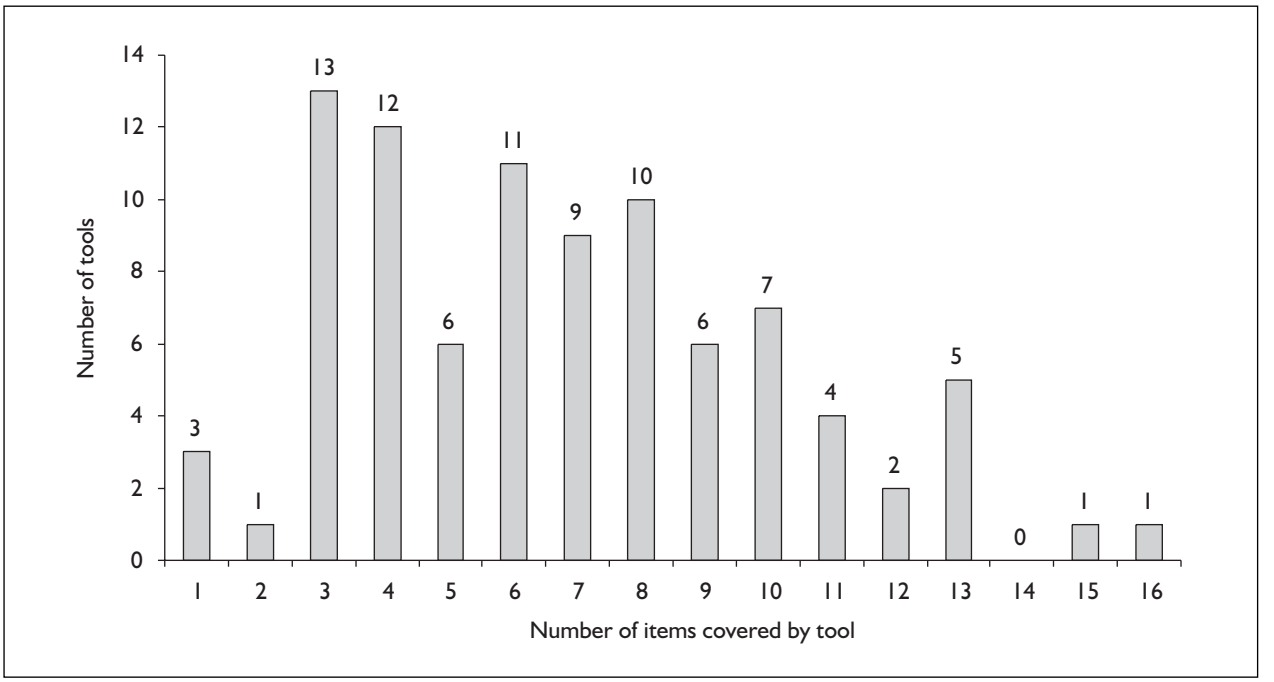

FIGURE 4 *Number of categories of bias covered by the quality assessment tools against number of tools*

range of categories covered was 1–16, with a mean of 6.8. This is illustrated in *Figure 4*, which shows the number of items covered by the tool against the number of tools.

TABLE 22 *Proportion of quality assessment tools that covered each item together with the classification of each item*

Category of bias	Source of bias	Proportion of tools (%)	Classification
Spectrum composition	Spectrum composition	64	II
	Inclusion criteria	12	IV
	Population recruitment	58	II
	Disease prevalence/severity	10	IV
Index test and reference standard			
Selection and execution	Absent or inappropriate reference standard	64	II
	Change in technology of index test	2	IV
	Disease progression bias	4	IV
	Test execution	40	III
	Reference execution	11	IV
	Verification bias	63	II
	Incorporation bias	8	IV
	Normal defined	38	III
	Treatment paradox	5	IV
Interpretation	Review bias	87	I
	Clinical review bias	11	IV
	Observer/instrument variation	35	III
Data presentation	Appropriate results	35	III
	Precision (sample size, variation by chance)	13	IV
	Inappropriate handling of uninterpretable test results	18	IV
	Post hoc choice of threshold value	3	IV
	Dropouts	7	IV
	Subgroups	9	IV
	Data table	12	IV
	Utility of test	16	IV
Research planning	Sample size	33	III
	Objectives	7	IV
	Protocol	5	IV

Levels of evidence for each classification of bias

Using the classification discussed in the section on Data synthesis (p. 39), each quality item was classified from I to IV according to the proportion of tools that included each item. A classification of I indicates that over 75% of tools included this quality item; a score of IV indicates that less than 25% of tools included this item. The classification of each source of bias is shown in *Table 22*.

Only one quality item, review bias, was rated as level I. Four items, spectrum composition, population recruitment, the use of an appropriate reference standard and verification bias, were rated as II, that is, they were included in more than half the quality assessment tools. Five items, observer/instrument variability, description of test execution, appropriate definition of a 'normal' test result, appropriate sample size and the presentation of appropriate results, were classified as III. All other items were rated as level IV.

Discussion

This review has provided a comprehensive overview of the quality assessment tools currently used to assess studies of diagnostic test accuracy. A systematic approach was adopted to identify existing tools and those that met inclusion criteria were classified in a systematic manner. A large number of tools (91) was identified. It is felt that it is unlikely that many existing quality assessment tools will have been missed by the searches, and if any were missed, that they would have had any major effect on the results of the review. The tools have been summarised according to how they were developed, the purpose for which they were developed and the quality items that they address.

The main limitation of the identified quality assessment tools was the lack of details on how the tools were developed. Only two of the 91 tools included in the review provided details of tool development. None of the tools appeared to be

validated. A large proportion of the tools identified (59%) were developed specifically within the context of systematic reviews or methodological reviews of diagnostic test studies to assess the quality of studies included in the review. The remaining tools were either original tools or modified versions of existing tools. These tools were developed either to assess study quality, as guides for interpreting the quality of published studies or as guides for reporting studies of diagnostic accuracy. Although these all relate to assessing study quality, the differences in the aims of the tools may be reflected by differences in items included in the tools.

Most tools used a 'checklist' type quality assessment (67%) and a further 12 used a scoring system to produce an overall quality score. Checklists have the advantage that different aspects of quality can be examined individually, while a quality score can be used to give a quick overview of the quality of each study. A small proportion of the tools used a level of evidence approach. This has the advantage that several items can be incorporated into the quality assessment to group studies into different levels, giving a quick impression of the general quality of each study. However, such tools can only incorporate a limited number of quality items and the effects of individual quality items cannot be assessed. In general, checklists are preferable to levels of evidence as these provide more information on the quality of the individual study by assessing studies for a number of different items.

There was a large variation in the number of quality assessment items addressed. Out of the 26 possible items the most items included on any one tool was 16, and some tools only looked at one item of quality. Different tools included different items, showing that there is disagreement regarding which features of quality are important in studies of diagnostic accuracy. Only one item, review bias, was included in over 75% of tools. Four items were included in 50–75% of tools: spectrum composition, population recruitment, the use of an appropriate reference standard and verification bias. Other items were included in less than 50% of tools. Each of the 27 items included in the list of possible items was included in at least one tool. Very few of the studies included topic-specific items. None of the original or modified tools included topic-specific items, even though several of these were developed for specific topic areas. Only a very small proportion of the tools developed specifically to assess the quality of studies included in reviews contained topic-specific items. These tools were all developed for very specific topic areas and none of the topic-specific items could be generalised to studies in similar areas.

The 27 items included in the list of possible items relate both to the internal validity and external validity of a study and to the quality of the reporting. Internal validity relates to whether estimates of diagnostic accuracy have been biased as a result of factors related to the design and conduct of the study. Possible sources of bias that may affect internal validity include population recruitment, absent or inappropriate reference standard, disease progression bias, verification bias, incorporation bias, treatment paradox, review bias, clinical review bias, inappropriate handling of uninterpretable test results and post hoc choice of threshold value. The external validity is related to the extent to which the results of a study can be applied to patients in practice. Factors included in the 27 items that may affect the external validity of a study include spectrum composition, disease prevalence/severity, change in technology of index test, test execution, reference standard execution, definition of a normal test result and observer/instrument variation. All other items included in the list of possible items relate to the quality of reporting rather than to the actual validity of the study.

Items that were classified as level I or II (spectrum composition, population recruitment, absent or inappropriate reference standard, verification bias and review bias) all relate to internal validity, with the exception of spectrum composition, which relates to external validity. Most of the items relating to the quality of reporting were rated as level IV, that is, they were included in less than 25% of tools. The only exceptions to this were the presentation of appropriate results and the use of an appropriate sample size; these were both rated as level III and were included in 35% and 33% of tools, respectively. This suggests that items relating to internal validity are generally considered the most important factors that need to be included in quality assessment tools of diagnostic accuracy. Theoretically, this would seem logical, as these are the items that are most likely to produce biased estimates of test performance. As external validity relates to the generalisability of results, aspects of this form of validity will be more important to consider when assessing whether the results of a study can be generalised to different settings. These factors are more likely to be included in checklists developed to interpret the results of a study of diagnostic accuracy. In general, factors

related to the reporting of study results were included in fewer scales. Such items will not directly bias estimates of test performance, but may provide an overall impression of the quality of a study. They will be more important for tools developed as guides for reporting studies than for those that were developed specifically to assess study quality.

A possible reason for the lack of agreement between tools on which items should be included in the tool and the poor coverage of quality items by these tools is the lack of empirical evidence regarding which sources of bias are likely to affect estimates of test performance. Compared to quality indicators for RCTs, very few similar studies have been conducted in the area of diagnostic studies. Further work, such as the study by Lijmer and colleagues,[33] is needed to provide more empirical evidence of the effects of different sources of bias on test performance. A similar project, replicating this study on a larger subset of meta-analyses using more sophisticated methods of analysis, is currently being undertaken by researchers at the Department of Clinical Epidemiology and Biostatistics at the University of Amsterdam. This information should be considered in the development of future quality assessment tools that should be developed in a systematic manner and validated empirically.

Conclusions

Existing tools used to assess the quality of studies of diagnostic accuracy suffer from a number of weaknesses. The main problem is the lack of details on scale development, and none of the tools was reported to be validated in any way. There is also a lack of agreement between tools regarding which items should be included in the tool. Only one item, the avoidance of review bias, was included in more than 75% of tools. A further four items were each included in 50–75% of tools: spectrum composition, population recruitment, absent or inappropriate reference standard and verification bias. Other items were included in less than 50% of tools. There is a need for a validated tool for the quality assessment of studies of diagnostic accuracy.

Chapter 8

Objective 4: Develop a new evidence-based assessment tool for the quality assessment of diagnostic studies

The results of the systematic reviews were used to develop a new evidence-based assessment tool. This followed the approach suggested by Streiner and Norman[226] in 'Health measurement scales: a practical guide to their development and use', which was also adopted by Jadad and colleagues[227] to establish a scale for assessing the quality of randomised controlled studies. The procedure involves the following steps:[227]

1. Preliminary conceptual decisions
2. Item generation and assessment of face validity
3. Field trials to assess consistency and construct validity
4. Generation of a refined instrument.

Preliminary conceptual decisions

For the purposes of this project 'quality' is defined as concerned with the internal and external validity of a study. This is, 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.

The aims of the new assessment tool were to:

- assess the scientific quality of a diagnostic study in generic terms (relevant to all diagnostic studies)
- allow consistent and reliable assessment of quality by raters with different backgrounds.

This project only produced the generic section of the quality assessment tool. Work on the topic-specific items will continue after this project has finished.

Based on the results of Chapter 6, it was decided that the quality assessment tool needed to have the potential to allow quality to be discussed narratively, be reported in a tabular summary, be used as recommendations for future research, be used to conduct sensitivity or regression analyses and be used as criteria for inclusion in the review or a primary analysis. The resulting implication for the development of the tool is that some distinction needs to be made between high- and low-quality studies. This may be done by producing a quality score, developing a levels of evidence approach or highlighting criteria that are more likely to lead to bias (component analysis). There are advantages and disadvantages associated with all these methods.

The issue of whether or not to use quality scores is the topic of ongoing debate in the field of systematic reviews of therapeutic trials.[228–234] Many of the issues raised in these discussions are equally relevant to the field of diagnostics. In calculating summary quality scores, the weight given to each item, which has been objectively rated, is determined subjectively and differs according to the quality scale used. The fact that the importance of individual items and the direction of potential biases associated with these items may vary according to the context in which they are applied is ignored.[234,235] The application of quality scales, with no consideration of the individual quality items, may therefore dilute or entirely miss potential associations.[236] It has also been shown that different quality scales produce very different indications of the quality of a study.[235] For these reasons it was decided not to incorporate a quality score into the quality assessment tool.

Another approach to the quality assessment of studies involves a 'levels of evidence' approach. Levels of evidence are slightly different: studies are assigned a level or grade if they fulfil a predefined set of items. There are usually several different levels, each with a different set of items that have to be met for a study to reach each level, with higher quality studies having to meet a more rigid set of items. For example, a high-quality study may be considered level 1, whereas a very poor-quality study which fails to meet any of the predefined items may be considered level 4. One of the problems associated with a levels of evidence approach is that each level incorporates

TABLE 23 *Summary of evidence provided for each source of bias from Chapters 5 and 7*

Category of bias	Bias	Chapter 5: Evidence of effect of bias (no. of studies)			Chapter 7: Classification[a]
		Empirical	Theoretical	No evidence	
Spectrum composition	Variation by clinical and demographic subgroups (spectrum composition)	14	0	1	II
	Inclusion criteria				IV
	Distorted selection of participants	3	0	2	II
	Disease prevalence/severity	8	1	0	IV
Index test and reference standard					
Selection and execution	Absent or inappropriate reference standard	4	4	0	II
	Change in technology of index test	1	0	1	IV
	Disease progression bias	0	0	1	IV
	Difference in test protocol (test execution)	1	0	1	III
	Difference in test protocol (reference execution)				IV
	Partial verification bias	17	3	3	II
	Differential verification bias	2	0	0	
	Incorporation bias	0	0	0	IV
	Normal defined				III
	Treatment paradox	0	0	0	IV
Interpretation	Review bias	4	0	1	I
	Clinical review bias	7	0	1	IV
	Observer/instrument variation	8	0	0	III
Analysis	Appropriate results				III
	Precision (sample size, variation by chance)	0	0	0	IV
	Inappropriate handling of uninterpretable/indeterminate/ intermediate test results	0	0	2	IV
	Post hoc choice of threshold value	0	0	0	IV
	Dropouts	0	0	0	IV
	Subgroups				IV
	Data table				IV
	Utility of test				IV
Research planning	Sample size				III
	Objectives				IV
	Protocol				IV

[a] Classification: I, included in >75% of tools; II, included in 50–75% of tools; III, included in 25–50% of tools; IV, included in <25% of tools.

several different quality items and so it is not possible to assess which of the individual quality items a study fulfils.

Checklists have the advantage that different aspects of quality can be examined individually and do not suffer from the problems associated with quality scores and levels of evidence. Therefore, it was decided that component analysis is the best approach to incorporate quality into a systematic review of diagnostic studies. The quality tool was developed taking this into consideration.

Item generation and assessment of item face validity

Item generation

An initial list of possible items for inclusion in the quality assessment tool was developed incorporating the results of the systematic reviews. The sources of bias set out in Chapter 4 were included in the initial list of items. The evidence provided from Chapters 5 and 7 was summarised for each source of bias (*Table 23*). The results from Chapter 5 were summarised according to the

number of studies providing empirical, theoretical or no evidence of bias. The results from Chapter 7 were summarised according to the proportion of studies that included each form of bias in the checklist. The results from Chapter 6 did not contribute directly to the checklist. These were used to provide an indication of what the checklists would be required to do, by identifying how existing systematic reviews of diagnostic tests had incorporated quality assessment into the reviews, rather than identifying items for inclusion in the checklist.

Each item was phrased as a question and sufficient detail was provided on the meaning of the item so that anyone assessing a study for quality would be able to apply the quality item to that study.

Assessment of face validity

A Delphi procedure was used to assess the face validity of the items included in the initial list of items and hence to select items for inclusion on the quality assessment tool. Assessment of face validity involves determining whether, on the face of it, the instrument appears to be assessing the desired qualities.[226] It is a qualitative process, involving the application of common sense using a global approach in which decisions are made using implicit but unspecified variables and criteria, with no statistical tactics.[237] It is, therefore, a subjective process and empirical approaches are rarely used.[226]

Delphi procedures aim to obtain the most reliable consensus amongst a group of experts by a series of questionnaires interspersed with controlled feedback.[238] The technique involves the recruitment of experts in a particular field and repeated questioning of each group member, using sequential questionnaires. A statistical summary of group responses is prepared following each round of questions. This is used in developing the next round of questions, and is issued as feedback so that individuals may revise their views through awareness of overall responses, rather than through pressure from individuals.[239] There are four distinguishing features of a Delphi technique:[238] anonymity (although members of the group may be identified their answers will be anonymous), iteration (the procedure involves several rounds), controlled feedback (the results of each round are analysed separately and responses fed back to members of the Delphi panel), and statistical group response (expression of the degree of consensus of the group on a particular issue).

Delphi procedure

Delphi panel members

As the area of diagnostic accuracy studies is a specialised area, it was decided to include a small number of experts in the area on the panel, rather than to include a larger number of participants who may have had a more limited knowledge of the area. Eleven experts in the area were contacted and asked to become panel members for the Delphi procedure. Two declined to take part in the procedure, eight completed all questionnaires and one completed only round 3 of the procedure. Details of panel members, all of whom were also members of the advisory group, are provided in Appendix 5.

Round 1

The initial list of 28 possible items for inclusion in the quality assessment tool, divided into four categories as above, was sent to all panel members. The aim was to collect information on each member of the group's opinion regarding the importance of each item. To help panel members in their decision-making, the evidence from Chapters 5 and 7 was summarised (as above) for each item. The aims of the quality assessment tool and its desired features were presented. Members of the panel were asked to rate each item for inclusion in the quality assessment tool according to a five-point Likert scale (strongly disagree, moderately disagree, neutral, moderately agree, strongly agree). Panel members were given the opportunity to comment on any of the items included in the tool, to suggest possible rephrasing of questions and to highlight any items that may have been missed off the initial list of items.

The results of round 1 are summarised in *Table 24*. Eight of the nine people who agreed to take part in the procedure returned completed questionnaires. The ninth panel member did not have time to take part in this round.

Round 2

The results of round 1 were used to select items for which there were high levels of agreement for inclusion or exclusion from the final quality assessment tool. All categories/items rated as 'strongly agree' by at least six of the eight panel members were selected for inclusion in the tool. Categories/items that were not rated as 'strongly agree' by at least one panel member were excluded. Based on the results of round 1, six items were selected for inclusion, one item was removed from the tool and the remaining items were rerated as part of round 2.

TABLE 24 *Results from the Delphi procedure round 1*

Category/Item		Likert score[a]				
		1	2	3	4	5
C1	*Spectrum composition*[b]	0	0	1	1	6
11	Was the spectrum of patients described in the paper and was it chosen adequately?	0	1	0	1	5
12	Were selection criteria described clearly?	0	0	3	1	3
13	Was the method of population recruitment consecutive?	0	0	4	2	1
14a	Was the setting of the study relevant?	0	1	2	1	3
14b	Was disease prevalence and severity reported?	0	1	0	2	4
C2a	*Index test and reference standard: selection and execution*	0	0	0	1	7
11	In light of current technology, was the reference standard chosen appropriate to verify test results?	0	0	0	1	7
12	Is it possible that a change in the technology of the index test has occurred since this paper was published?	0	2	1	4	1
13	Was there an abnormally long period between the performance of the test under evaluation and the confirmation of the diagnosis with the reference standard?	0	1	0	5	2
14	Was the execution of the index test described in sufficient detail to permit replication of the test?	0	0	2	1	5
15	Was the execution of the reference standard described in sufficient detail to permit replication of the test?	0	0	2	1	5
16	Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?	0	0	2	0	6
17	Did all patients receive the same reference standard regardless of the index test result?	0	0	1	2	5
18	Were the results of the index test incorporated in the results of the reference standard?	1	0	1	3	3
19	Was the cut-off value prespecified or acceptable in light of previous research?	0	1	3	1	3
110	Was treatment started based on the knowledge of the index test results before the reference standard was applied?	0	1	4	0	3
C2b	*Index test and reference standard: interpretation*	0	0	0	0	8
11a	Were the index test results interpreted blind to the results of the reference standard?	0	0	0	0	8
11b	Were the reference standard results interpreted blind to the results of the index test?	0	0	0	1	7
12	Were clinical data available when test results were interpreted?	0	1	0	1	6
13	Were data presented on observer or instrument variation that could have affected the estimates of test performance?	0	1	2	1	4
C3	*Analysis*	0	0	0	1	7
11	Were appropriate results presented and were these calculated appropriately?	1	0	3	0	4
12	Was a measure of precision of the results presented?	1	0	2	2	3
13	Were uninterpretable/indeterminate/intermediate results reported and included in the results?	0	1	0	1	6
14	Was the threshold value specified retrospectively based on analysis of the results?	0	0	4	2	2
15	Were reasons for dropout from the study reported?	0	0	1	3	4
16	Were subgroup analyses prespecified and clinically relevant?	0	0	4	2	2
17	Were results presented in a 2 $\times$ 2 table?	0	2	3	0	3
18	Was any indication of the utility of the test provided?	0	0	7	1	0
C4	*Research planning*[c]	0	1	4	1	2
11	Was an appropriate sample size calculation performed?	0	1	0	1	1
12	Were study objectives clearly reported?	0	0	0	0	3
13	Was there any evidence that a study protocol had been developed before the study started?	0	1	0	0	2

[a] Likert score: 1, strongly disagree: 2, moderately disagree: 3, neutral: 4, moderately agree: 5, strongly agree.
[b] The category 'Spectrum bias' was only rated by seven panel members, as one member rated this category as neutral.
[c] The category 'Research planning' was only rated by three panel members, as five panel members rated this category as neutral or less.
C, category; I, item within a category.

Items selected for inclusion were:

- appropriate selection of patient spectrum
- appropriate reference standard
- absence of partial verification bias
- absence of review bias (both test and diagnostic)
- clinical review bias
- reporting of uninterpretable/indeterminate/intermediate results.

The item removed from the tool was:

- test utility.

Panel members made a number of suggestions regarding rephrasing of items. These were considered and changes were made where appropriate. Based on some of the comments received two additional items were added, one to the category 'Spectrum composition' and the second to the category 'Analysis'. The first item was 'What was the study design?' and the second was 'Were sufficient data provided to include the study in a systematic review?' These items were rated for inclusion in the tool as part of round 2.

Items selected for inclusion or exclusion from the final quality assessment tool were not rated as part of round 2. The results from the first round were reported to provide panel members with a summary of the responses of all panel members. This was provided in two sections: a summary of the comments for each category and item, and a summary of the ratings for each category and item.

Details were provided to panel members on how decisions were reached regarding which items to include in the final quality assessment tool. Several other decisions were also made, including how to handle missing responses, rephrasing of items and definition of adequate/appropriate/abnormally; these were also reported together with the justification for the decisions.

For the round 2 questionnaire, rather than rating each item on the five-point Likert scale, panel members were asked to indicate whether they thought that a category or item should be included or excluded from the quality assessment tool. They were asked to consider the results from round 1, the comments from round 1 and the evidence provided for each item when making this decision. In addition, panel members were asked to answer yes or no to the following questions.

- Would you like to see a number of 'key items' highlighted in the quality assessment tool?
- Do you endorse the Delphi procedure so far? If no, please give details of the aspects of the procedure that you do not support and list any suggestions you have for how the procedure could be improved.
- As part of the third round, instructions on how to complete the quality assessment will be provided to you. As we do not want to ask you to invest too much time, the instructions will be drawn up by the steering group. In the third round you will only be asked whether you support the instructions and, if not, what you would like to change. Do you agree with this procedure?

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

The results of round 2 are summarised in *Tables 25* and *26*. Of the nine people invited to take part in round 2, eight returned completed questionnaires.

Round 3

The results of round 2 were used further to select items for inclusion or exclusion in the quality assessment tool. All categories rated as include by at least seven out of eight of the panel members were selected for inclusion in the tool. Items scored 'include' by six out of eight of the panel members were rerated as part of round 3. All other items were excluded. Based on the results of this round, a further four items were selected for inclusion in the tool:

- absence of disease progression bias
- absence of differential verification bias
- absence of incorporation bias
- reporting of study withdrawals.

Panel members did not achieve consensus for a further five items; these were rerated as part of round 3:

- reporting of selection criteria
- reporting of disease severity
- description of index test execution
- description of reference standard execution
- independent derivation of cut-off points.

All other items, including the new items added based on feedback from round 1, were excluded from the process.

Since none of the panel members was in favour of highlighting a number of key items in the quality

TABLE 25 *Results from the Delphi procedure round 2: items for inclusion in the tool*

Category/Item		Decision		
		Include	Exclude	?
Included items				
C1	*Spectrum composition*			
I1	Was the spectrum of patients selected appropriately?			
C2a	*Index test and reference standard: selection and execution*			
I1	Is the reference standard likely to classify correctly the target condition?			
I6	Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?			
C2b	*Index test and reference standard: interpretation*			
I1a	Were the index test results interpreted without knowledge of the results of the reference standard?			
I1b	Were the reference standard results interpreted without knowledge of the results of the index test?			
I2	Were clinical data available when test results were interpreted?			
C3	*Analysis*			
I3	Were uninterpretable/indeterminate/intermediate results reported?			
Items rerated during round 2				
C1	*Spectrum composition*			
New	What was the study design?	5	3	0
I2	Were selection criteria clearly described?	6	1	1
I3	Was a random or consecutive sample of patients included in the study?	5	2	1
I4a	Was the setting of the study relevant?	0	8	0
I4b	Was disease prevalence reported?	5	3	0
I4c	Was disease severity reported?	6	1	1
C2a	*Index test and reference standard: selection and execution*			
I2	Is it possible that a change in the technology of the index test has occurred since this paper was published?	2	6	0
I3	Is the period between reference standard and index test short enough to be reasonably sure that disease status did not change between the two tests?	7	1	0
I4	Was the execution of the index test described in sufficient detail to permit replication of the test?	6	2	0
I5	Was the execution of the reference test described in sufficient detail to permit replication of the test?	6	2	0
I7	Did patients receive the same reference standard regardless of the index test result?	8	0	0
I8	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	8	0	0
I9	Was the definition of a 'normal' test result reported?	5	3	0
I10	Was treatment started based on the knowledge of the index test results before the reference standard was applied (treatment paradox)?	5	3	0
C3	*Analysis*			
New	Were sufficient data provided to include the study in a systematic review?	2	7	0
I1a	Were appropriate results presented?	3	5	0
I1b	Were results calculated appropriately?	3	5	0
I2	Was a measure of precision of the results presented?	4	4	0
I3a	Were data presented on observer variation?	3	5	0
I3b	Were data presented on instrument variation?	3	5	0
I4	Were threshold values used in interpreting results derived independently to the results of the current study?	6	2	0
I5	Were both numbers and reasons for dropout from the study reported?	8	0	0
I7	Were sufficient results presented to calculate an $n \times n$ (e.g. 2×2) data table?	4	4	0
I6a	Were subgroup analyses prespecified?	1	7	0
I6b	Were subgroup analyses clinically relevant?	2	6	0

continued

TABLE 25 *Results from the Delphi procedure round 2: items for inclusion in the tool (cont'd)*

Category/Item		Decision		
		Include	Exclude	?
C4	*Research planning*[a]	4	4	0
I1	Were sufficient participants included in the study?	0	4	0
I2	Were study objectives clearly reported?	4	0	0
I3	Was there any evidence that a study protocol had been developed before the study started?	3	1	0
Excluded items				
I8	Was any indication of the utility of the test provided?			

Shaded areas indicate items that were 'included' or 'excluded' from the quality assessment tool based on the results of round 1 and were therefore not rerated as part of round 2.
[a] The category 'Research planning' was only rated by four panel members, as two panel members rated this category as 'exclude'.

TABLE 26 *Results from the Delphi procedure round 2: additional items*

	Response		
	Yes	No	Unclear
Question			
1. Would you like to see a number of 'key items' highlighted in the quality assessment tool?	0	8	0
2. Do you endorse the Delphi procedure so far?	5	1	2
3. Procedure for instructions on completing tool	8	0	0
Validation step			
1. Piloting by three raters	8	0	0
2. Exclusion of frequency of endorsement step. Do you agree with this?	7	0	1
3. Assessment of consistency and reliability of the instrument	6	0	2
4. The instrument will be adjusted based on the outcome of the above steps	7	0	1
5. Regression analysis	5	1	2
6. The tool will be piloted in a number of diagnostic reviews	8	0	0

assessment tool, this approach was not followed. At this stage, five of the panel members reported that they endorsed the Delphi procedure so far, one did not and two were unclear. The member who did not endorse the Delphi procedure stated that "I fundamentally believe that it is not possible to develop a reliable discriminatory diagnostic assessment tool that will apply to all, or even the majority of diagnostic test studies." One of the comments from a panel member who was 'unclear' also related to the problem of producing a quality assessment tool that applies to all diagnostic accuracy studies. The other related to the process used to derive the initial list of items and the problems of suggesting additional items. All panel members agreed to let the steering group produce the background document to accompany the tool. The feedback suggested that there was some confusion regarding the proposed validation methods. These were clarified and rerated as part of round 3.

The results from round 2 were reported to provide panel members with a summary of the responses of all panel members. This was provided in two sections: a summary of the comments for each category and item, and a summary of the ratings for each category and item. Details were provided to panel members on how decisions were reached regarding which items to include in the final quality assessment tool.

All items selected for inclusion in the tool at this stage were presented and panel members were asked to indicate whether they agreed with the proposed phrasing of the items and, if not, to suggest alternative phrasings. As for the round 2 questionnaire, panel members were asked to indicate whether they thought that each item to be rerated should be included or excluded from the quality assessment tool. They were asked to consider the results from round 2, the comments from round 2 and the evidence provided for each

item when making this decision. Based on the feedback from round 2 several additional items were proposed. Panel members were asked to rate these as yes or no according to whether they thought that these should be included in the tool.

Some additional questions were included for this round. A scoring system was proposed and panel members were asked to indicate whether they agreed with this system. Based on the results of round 2, there appeared to be some confusion regarding the proposed validation methods. Further details of the proposed methods were therefore presented and panel members were again asked to indicate whether they agreed with these methods. The aims of the quality assessment tool were highlighted and panel members were asked whether, taking these into consideration, they endorsed the Delphi procedure. Members were also asked whether they used the evidence provided from the reviews and the feedback from previous rounds in their decisions of which items to select for inclusion in the tool, and if they did not use this information to explain why not. Lastly, panel members were asked whether they would like to see the development of topic- and design-specific items in addition to the generic section of the tool. If they answered yes to these questions they were asked whether they would like to see the development of these items through a further Delphi procedure and, if so, whether they would like to be a member of the panel for this procedure. It was decided to name the tool the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool. A background document to accompany the QUADAS tool was produced for the items selected for inclusion in the tool up to this point. Panel members were asked to comment on this.

Based on the comments from round 2, four following additional items were included in the round 3 questionnaire; these are detailed in *Table 27*.

The results of round 3 are presented in *Tables 27* and *28*. All nine panel members invited to take part in round 3 returned completed questionnaires.

Round 4

Agreement was reached on items to be included in the tool following the results of round 3, and so round 4 was the final round of the Delphi procedure. The results of round 3 were used to select the final items for inclusion or exclusion in the quality assessment tool. All categories rated as include by at least seven out of nine of the panel members were selected for inclusion in the tool. All other items were removed from the tool.

Three of the five items rerated as part of round 3 were selected for inclusion. These were:

- reporting of selection criteria
- description of index test execution
- description of reference standard execution.

The other two items and the additional items rated as part of this round were not included in the tool. Comments regarding rephrasing of items were considered and items were rephrased taking these into account.

Each of the proposed validation steps were approved by at least seven out of nine of the panel members and so these methods will be used to validate the tool. More than half of the panel members indicated that they would like to see the development of design- and topic-specific criteria, and of these four stated that they would like to see this done via a Delphi procedure. The development of these elements will therefore take place after the generic section of the tool has been validated.

As for the earlier rounds, the results from round 3 were reported to provide panel members with a summary of the responses of all panel members. This was provided in two sections: a summary of the comments for each category and item, and a summary of the ratings for each category and item. Details were provided to panel members on how decisions were reached regarding which items to include in the final quality assessment tool.

The final version of the QUADAS tool consists of 14 items. The QUADAS tool was presented; the background document was adjusted based on the feedback from round 3 and the updated version presented. Panel members were given a final chance to comment on any features of the QUADAS tool or background document with which they were unhappy.

Endorsement of the Delphi procedure and use of the evidence supplied

All but one of the panel members stated that they endorsed the Delphi procedure. This member remained unclear as to whether he endorsed the procedure and stated that "all my reservations still apply". These reservations relate to comments made in earlier rounds regarding the problems of developing a quality assessment tool that can be applied to all studies of diagnostic accuracy. Seven

TABLE 27 *Results from the Delphi procedure round 3: rating of items to be included*

Items to be rerated	Include	Exclude
I Should this item be included in the criteria list?		
1. Were selection criteria clearly defined?	7	2
2. Was disease severity reported?	4	5
3. Was the execution of the index test described in sufficient detail to permit replication of the test?	7	2
4. Was the execution of the reference standard described in sufficient detail to permit its replication?	7	2
5. Were threshold values used in interpreting results derived independently to the results of the current study?	3	6
Additional items	**Include**	**Exclude**
I Should this additional item be included in the criteria list?		
1. Are there other aspects of the design of this study which cause concern about whether or not it will correctly estimate test accuracy?	3	6
2. Are there other aspects of the conduct of this study which cause concern about whether or not it will correctly estimate test accuracy?	4	5
3. Are there special issues concerning patient selection which might invalidate test results?	3	6
4. Are there special issues concerning the conduct of tests which might invalidate test results?	3	6

TABLE 28 *Results from the Delphi procedure round 3: additional items*

Question	Response		
	Yes	No	Unclear
1. Do you agree with the scoring proposed for the quality assessment tool?	9	0	0
2. Validation step			
Exclusion of frequency of endorsement step. Do you agree with this?	7	2	0
Assessment of consistency and reliability of the instrument	9	0	0
The instrument will be adjusted based on the outcome of the above steps	9	0	0
Regression analysis	8	1	0
3. Given the comments from the steering group at the bottom of page 10, do you endorse the Delphi procedure so far?	8	0	1
4. Did you use the evidence provided from the systematic reviews to help make decisions on which items to include in the quality assessment tool?	7	2	0
5. Did you use the evidence provided from the feedback from round 1 to help make decisions on which items to include in the quality assessment tool?	6	3	0
6. Would you like to see the development of topic-specific items in addition to the generic quality assessment tool?	5	4	0
7. Would you like to see the development of design-specific items in addition to the generic quality assessment tool?	5	4	0
8. If you would like to see the development of topic- and/or design-specific items would like to see this done via a Delphi procedure?	4	1	0
9. If you would like to see the development of topic- and/or design-specific items via a Delphi procedure, would you like to be part of the Delphi panel?	4	0	0

of the panel members reported using the evidence provided from the systematic reviews to help in their decisions of which items to include in the QUADAS tool. Of the two that did not use the evidence, one stated that they were too busy and the other stated that there was "no new stuff inside". Seven of the panel members reported using the feedback from earlier rounds of the Delphi procedure. Of the two that did not, one stated that they were "not seeking conformity with other respondents" and the other gave no explanation for not using the feedback. The two that did not use the feedback were different from the two that did not use the evidence provided by the reviews. These responses suggest that the evidence provided by the review contributed towards the production of the QUADAS tool.

The questionnaires for each round, including feedback from the previous rounds, are provided in Appendix 6. The QUADAS tool and the background document to accompany the tool are presented in Chapter 9.

Chapter 9

QUADAS background document

Background to the tool

Studies of diagnostic tests have commonly addressed one of two main objectives, and the chosen study methodology is likely to reflect this. The first, and traditionally the most common aim of diagnostic test evaluation, is to establish the diagnostic accuracy of the test. The second objective of diagnostic research is to evaluate the impact of one or more diagnostic strategies on therapy decisions and/or patient outcomes. Diagnostic impact tends to be assessed in either RCTs or non-experimental comparative studies, and so is affected by different quality issues. The QUADAS tool has been developed to assess the quality of studies of diagnostic accuracy.

Diagnostic accuracy studies follow a similar basic structure. They aim to determine how good a particular test, the index test, is at detecting the target condition. A series of patients receives the test (or tests) of interest, known as the index test(s), and also a reference standard. The results of the index test(s) or interpretations thereof are then compared to the results of the reference standard. The reference standard should be the best available method to determine whether or not the patient has the condition being tested for. It may be a single test, clinical follow-up or a combination of tests. Both the terms 'test' and 'condition' are interpreted in a very broad sense. The term 'test' is used to refer to any procedure used to gather information on the health status of an individual. This can include laboratory tests, surgery, clinical examination, imaging tests, questionnaires and pathology. Similarly, 'condition' can be used to define any health status, including the presence of disease (e.g. influenza, alcoholism, depression, cancer), pregnancy or different stages of disease (e.g. an exacerbation of multiple sclerosis).

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

Diagnostic studies have several unique features in terms of quality, such as verification and spectrum bias, that are not addressed by the traditional approach to evaluating controlled trials (which has focused on randomisation, allocation concealment and blinded outcome assessment). If quality is not assessed appropriately then false conclusions may be drawn from biased studies. It is therefore essential that the quality of individual studies included in a systematic review is assessed, in terms of the lack of applicability and the potential for bias. This is the first tool for the assessment of the quality of diagnostic accuracy studies which has been systematically developed and is evidence based. Full details on the development of the quality assessment tool, together with the evidence on which it is based, are available elsewhere.

This background document aims to introduce the QUADAS tool and to facilitate its use. It starts with the presentation of the tool, followed by a description of the meaning of each item included in the QUADAS tool, suggestions for situations in which it may not be appropriate to assess studies for a particular item, and guides on how to score each of the items.

Aims of the tool

The tool has been developed to assess the quality of diagnostic accuracy studies included in systematic reviews so that appropriate conclusions can be drawn in light of the potential biases. It is anticipated that the tool will be used:

- as criteria for including/excluding studies in a review
- as criteria for including/excluding studies in primary analyses
- to conduct sensitivity/subgroup analysis stratified according to quality
- as individual items in meta-regression analyses
- to make recommendations for future research.

The tool does not incorporate a quality score. Instead, it is structured as a list of 14 questions that should each be answered 'yes', 'no' or 'unclear'.

The quality assessment tool

TABLE 29 *The quality assessment 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 classify the target condition correctly?			
4. Is the 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?			

Explanation of items included in the quality assessment tool and guide to scoring items

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

What is meant by this item

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

Situations in which this item does not apply

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

How to score this item

Studies should score 'yes' for this item if you believe, based on the information reported or obtained from the study's authors, that the spectrum of patients included in the study was representative of those in whom the test will be used in practice. The judgement should be based on both the method of recruitment and the characteristics of those recruited. Studies that recruit a group of healthy controls and a group known to have the target disorder will be coded as 'no' on this item in nearly all circumstances. Reviewers should prespecify in the protocol of the review what spectrum of patients would be acceptable, taking factors such as disease prevalence and severity, age and gender into account. If you think that the population studied does not fit into what you specified as acceptable, the study should be scored as 'no'. If there is insufficient information available to make a judgement then it should be scored as 'unclear'.

2. Were selection criteria clearly described?

What is meant by this item

This refers to whether studies have provided a

clear definition of the criteria used as selection criteria for entry into the study.

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

How to score this item
If you think that all relevant information regarding how participants were selected for inclusion in the study has been provided then this item should be scored as 'yes'. If study selection criteria are not clearly reported then this item should be scored as 'no'. In situations where selection criteria are partially reported and you feel that you do not have enough information to score this item as 'yes', then it should be scored as 'unclear'.

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

What is meant by this item
The reference standard is the method used to determine the presence or absence of the target condition. To assess the diagnostic accuracy of the index test its results are compared with the results of the reference standard; subsequently, indicators of diagnostic accuracy can be calculated. The reference standard is therefore an important determinant of the diagnostic accuracy of a test. The reference standard may be obtained in many ways, including laboratory tests, imaging tests, function tests and pathology, but also clinical follow-up of participants. The decision regarding which reference standard to use depends on the definition of the target condition and the purpose of the study. If no single reference test is available, then careful clinical follow-up, a consensus between observers or the results of two or more combined tests may be used to determine the presence or absence of the target condition. Estimates of test performance are based on the assumption that the index test is being compared to a reference standard that is 100% sensitive and specific. If there are any disagreements between the reference standard and the index test then it is assumed that the index test is incorrect. Thus, from a theoretical point of view the choice of an appropriate reference standard is very important.

Situations in which this item does not apply
This item is relevant to all studies of diagnostic accuracy and should always be included in the quality assessment tool. The only exception would be if a particular reference standard were specified in the inclusion criteria; that is, to be included in the review a study may have to compare the index test to a specified reference standard.

How to score this item
If you believe that the reference standard is likely to classify the target condition correctly then this item should be scored 'yes'. Making a judgement as to the accuracy of the reference standard may not be straightforward. You may need experience of the topic area to know whether a test is an appropriate reference standard, or if a combination of tests is used you may have to consider carefully whether these were appropriate. If you do not think that the reference standard was likely to have classified the target condition correctly then this item should be scored as 'no'. If there is insufficient information to make a judgement then this should be scored as 'unclear'.

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

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

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

How to score this item
When to score this item as 'yes' is related to the target condition. For conditions that progress rapidly even a delay of several days may be important. For such conditions this item should be scored 'yes' if the delay between the performance of the index and reference standard is very short, a matter of hours or days. However, for chronic conditions disease status is unlikely to change in a week, a month, or even longer. In such conditions longer delays between performance of the index and reference standard may be scored as 'yes'. You

will have to make judgements regarding what is considered 'short enough'. You should think about this before starting work on a review, and define what you consider to be short enough for the specific topic area that you are reviewing. If you think the period between the performance of the index test and the reference standard was sufficiently long that disease status may have changed between the performance of the two tests then this item should be scored as 'no'. If insufficient information is provided this should be scored as 'unclear'.

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

What is meant by this item

Partial verification bias [also known as work-up bias, (primary) selection bias or sequential ordering bias] occurs when not all of the study group receive confirmation of the diagnosis by a reference standard. If the results of the index test influence the decision to perform the reference standard then biased estimates of test performance may arise. If patients are randomly selected to receive the reference standard the overall diagnostic performance of the test is, in theory, unchanged. In most cases, however, this selection is not random, possibly leading to biased estimates of the overall diagnostic accuracy.

Situations in which this item does not apply

Partial verification bias generally only occurs in diagnostic cohort studies in which patients are tested by the index test before the reference standard. If the test sequence is reversed, as it is in case–control designs, partial verification bias is generally not applicable. However, there may be exceptions to this. For example, in radiological rereading studies, scans are read at a later date by one or more radiologists, but the scans will usually have been obtained in regular clinical practice. If the study is limited to those with, for example, biopsy verification the index (radiological interpretations) could by influenced by the decision whether or not to biopsy, and verification bias may apply. In situations where the reference standard is assessed before the index test, you should first decide whether there is a possibility that verification bias could occur, and if not how to score this item. This may depend on how quality will be incorporated in the review. There are two options: either to score this item as 'yes', or to remove it from the quality assessment tool.

How to score this item

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

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

What is meant by this item

Differential verification bias occurs when some of the index test results are verified by a different reference standard. This is especially a problem if these reference standards differ in their definition of the target condition; for example, histopathology of the appendix and natural history for the detection of appendicitis. This usually occurs when patients testing positive on the index test receive a more accurate, often invasive, reference standard than those with negative test results. The link (correlation) between a particular (negative) test result and being verified by a less accurate reference standard will affect measures of test accuracy in a similar way as in partial verification, but less seriously.

Situations in which this item does not apply

Differential verification bias generally only occurs in diagnostic cohort studies in which all patients are tested by the index test before the reference standard. However, there may be situations in which this does not apply (see item 3). If the test sequence is reversed, as it is in case–control designs, partial verification bias is not applicable. In situations where the reference standard is assessed before the index test, you should decide how to score this item. This may depend on how quality will be incorporated in the review. There are two options: either to score this item as 'yes', or to remove it from the quality assessment tool.

How to score this item

If it is clear that patients received verification of their true disease status using the same reference standard then this item should be scored as 'yes'. If some patients received verification using a different reference standard than this item should be scored as 'no'. If this information is not reported by the study then it should be scored as 'unclear'.

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

What is meant by this item

When the result of the index test is used in establishing the final diagnosis, incorporation bias may occur. This incorporation will probably increase the amount of agreement between index test results and the outcome of the reference standard, and hence overestimate the various measures of diagnostic accuracy. It is important to note that knowledge of the results of the index test alone does not automatically mean that these results are incorporated in the reference standard. For example, a study investigating magnetic resonance imaging (MRI) for the diagnosis of multiple sclerosis could have a reference standard composed of clinical follow-up, cerebrospinal fluid analysis and MRI. In this case the index test forms part of the reference standard. If the same study used a reference standard of clinical follow-up and the results of the MRI were known when the clinical diagnosis was made but were not specifically included as part of the reference, then the index test does not form part of the reference standard.

Situations in which this item does not apply

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

How to score this item

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

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

What is meant by these items

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

Situations in which these items do not apply

These items are likely to apply in most situations.

How to score these items

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

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

What is meant by these items

This item is similar to blinding in intervention studies. Interpretation of the results of the index test may be influenced by knowledge of the results of the reference standard, and vice versa. This is known as review bias, and may lead to inflated measures of diagnostic accuracy. The extent to which this may affect test results will be related to the degree of subjectiveness in the interpretation of the test result. The more subjective the interpretation the more likely that the interpreter can be influenced by the results of the index test in interpreting the reference standard, and vice versa. It is therefore important to consider the topic area that you are reviewing and to determine whether the interpretation of the index test or reference standard could be influenced by knowledge of the results of the other test.

Situations in which these items do not apply

If, in the topic area that you are reviewing, the index test is always performed first then interpretation of the results of the index test will usually be without knowledge of the results of the reference standard. Similarly, if the reference standard is always performed first (e.g. in a diagnostic case–control study) then the results of

the reference standard will be interpreted without knowledge of the index test. However, in certain situations the results of both the index test and reference standard are blinded in both directions before being interpreted. In situations where one form of review bias does not apply there are two possibilities: either score the relevant item as 'yes' or remove this item from the list. If tests are entirely objective in their interpretation then test interpretation is not susceptible to review bias. In such situations review bias may not be a problem and these items can be omitted from the quality assessment tool. Another situation in which this form of bias may not apply is when test results are interpreted in an independent laboratory. In such situations it is unlikely that the person interpreting the test results will have knowledge of the results of the other test (either index test or reference standard).

How to score these items

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

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

What is meant by this item

The availability of information on clinical data during interpretation of test results may affect estimates of test performance. In this context clinical data are defined broadly to include any information relating to the patient obtained by direct observation, such as age, gender and symptoms. The knowledge of such factors can influence the diagnostic test result if the test involves an interpretative component. If clinical data will be available when the test is interpreted in practice then these should also be available when the test is evaluated. If, however, the index test is intended to replace other clinical tests then clinical data should not be available. It is therefore important to determine what information will be available when test results are interpreted in practice before assessing studies for this item.

Situations in which this item does not apply

If the interpretation of the index test is fully automated and involves no interpretation then this item may not be relevant and can be omitted from the quality assessment tool.

How to score this item

If clinical data would normally be available when the test is interpreted in practice and similar data were available when interpreting the index test in the study then this item should be scored as 'yes'. Similarly, if clinical data would not be available in practice and these data were not available when the index test results were interpreted then this item should be scored as 'yes'. If this is not the case then this item should be scored as 'no'. If this information is not reported by the study then it should be scored as 'unclear'.

13. Were uninterpretable/intermediate test results reported?

What is meant by this item

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

Situations in which this item does not apply

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

How to score this item

If it is clear that all test results, including uninterpretable/indeterminate/intermediate results, are reported then this item should be scored as 'yes'. If you think that such results occurred but have not been reported then this item should be scored as 'no'. If it is not clear whether all study results have been reported then this item should be scored as 'unclear'.

14. Were withdrawals from the study explained?

What is meant by this item

This occurs when patients withdraw from the study before the results of both the index test and reference standard are known. If patients lost to follow-up differ systematically from those who

remain, for whatever reason, then estimates of test performance may be biased.

Situations in which this item does not apply

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

How to score this item

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

Chapter 10

Discussion and proposals for further work

This project has produced a quality assessment tool to be used in systematic reviews of diagnostic accuracy studies. Through the various stages of the project the current lack of such a tool and the need for a systematically developed validated tool have been demonstrated.

The first section of the project (Chapter 5) reviewed the evidence for the types of bias that may affect diagnostic test evaluations. It provided an indication of the evidence available for the effects of each source of bias. The second section (Chapter 6) evaluated how quality assessment is currently incorporated into systematic reviews of diagnostic accuracy studies. In so doing, it highlighted the requirements of a quality assessment tool to be used to assess the quality of studies of diagnostic accuracy included in systematic reviews. In addition, this review highlighted the fact that not all systematic reviews of diagnostic accuracy studies perform a quality assessment and that those that do often fail to incorporate this appropriately in the review. It also showed that current reviews use a wide range of quality assessment tools, often developed specifically for the review by the review authors. The third section (Chapter 7) reviewed existing quality assessment tools for studies of diagnostic accuracy. This section fulfilled two functions: it highlighted the weaknesses of existing tools (mainly the lack of details on tool development and validation) and provided an overview of the items included in existing tools. The evidence from Chapters 5 and 7 contributed directly to the development of QUADAS. As part of the Delphi procedure panel members were provided with a summary of the evidence for the effects of bias for each item, together with an indication of the proportion of existing tools that included each item.

The QUADAS tool has therefore been developed in a systematic manner and is based on a review of existing evidence. This is the first time that a quality assessment tool for studies of diagnostic accuracy has been developed in this way.

The validation of the QUADAS tool was not performed as part of this project, but work to do this continues outside the project. The proposed steps for the validation of the tool are as follows.

1. The instrument will be piloted by three raters on a sample of published studies in order to identify any problems in clarity or application of the items. The items will then be reworded and instructions clarified if necessary.

2. The consistency or reliability can be measured by the degree to which different individuals agree on the scientific quality of a set of papers. A mixed group of raters consisting of researchers and clinicians will assess the same set of studies. The raters will be asked to assess the quality of the report independently, using the quality assessment tool and the background document, with no additional training on how to score the items. Intraclass correlation coefficients (ICCs) (for a more detailed description of these see Shrout and Fleiss[240]) and their 95% confidence intervals (CIs) will be used to measure the agreement between raters. Although any choice of cut-off will be arbitrary, the plan is to use the same cut-offs as used by Jadad and colleagues.[227] Items with ICCs greater than 0.50 will be considered to be sufficiently reliable, and those scoring greater than 0.65 represent a high level of agreement.

3. The tool will be piloted in a number of diagnostic reviews. Current projects planned include reviews of tests for tuberculosis, urinary tract infection in children, appendicitis, prediction of pre-eclampsia and prediction of preterm labour. Every reviewer piloting the review will be asked to complete a simple structured questionnaire that will gather information on how they used QUADAS and their opinions of it. Reviewers will also be asked to provide summaries of the results of their review and the quality assessment. Ideally, this will be in the form of 2×2 table data together with the results of the quality assessment, separately for each study included in the review. These data may be used as part of the analysis proposed in step 5.

4. The instrument will be adjusted based on the outcome of the above steps. Although the actual items included in the quality assessment tool will not be changed, the phrasing and

instructions accompanying each item may need to be adjusted if it appears that these are not being interpreted and applied as intended.

5. A regression analysis will be used to investigate associations between study characteristics and estimates of diagnostic accuracy in primary studies, as combined in existing systematic reviews. The methods used to conduct this analysis will be similar to the approach taken by Lijmer and colleagues.[33] A regression model adapted from the summary ROC curve developed for meta-analyses of diagnostic tests will be fitted to the data.[2] The logarithm of the diagnostic odds ratio (DOR) computed for a single study will be modelled as the dependent variable. Dependent variables for the intercept and slope of the curve will be fitted for each meta-analysis. Covariates for each methodological feature of the new assessment scale will be added simultaneously to this model. The resulting parameter estimates of the covariates can be interpreted after antilogarithm transformation as relative diagnostic odds ratios. They indicate the diagnostic performance of a test in studies failing to satisfy the methodological criterion, relative to its performance in studies with the corresponding feature. If the relative diagnostic odds ratio is larger than 1, studies not satisfying the criterion yield larger estimates of the diagnostic odds ratio than studies with this feature. This process will be carried out for several meta-analyses with relatively large numbers of included studies, for both diagnostic accuracy outcomes and therapeutic and/or patient outcomes. In addition to the analysis done by Lijmer and colleagues,[33] looking at the associations between characteristics and diagnostic odds ratios, the association of these characteristics with sensitivity and specificity will also be investigated.

The current QUADAS tool is a generic quality assessment tool. Proposed work will further develop the tool by adding topic- and design-specific items. These items will also be developed using a Delphi procedure. However, it is anticipated that several panels will be assembled to do this, including topic-specific experts for each of the topics covered.

The recent publication of the STARD document[241] stresses that the quality of reports of diagnostic accuracy study should be improved. QUADAS may also play a role in bringing about greater awareness regarding the important quality issues involved in diagnostic accuracy studies, helping to raise the standards of such studies.

In conclusion, this project has developed an evidence-based quality assessment tool for the assessment of studies of diagnostic accuracy included in systematic reviews. Further work to validate the tool continues beyond the scope of this project. The further development of the tool by the addition of design- and topic-specific criteria is proposed.

Acknowledgements

We would like to thank Kath Wright (Centre for Reviews and Dissemination) for carrying out literature searches and the advisory panel to the review for their help during various stages, including commenting on the protocol and draft report. We thank Dr Afina Glas for her support in the assessment of articles while reviewing the literature on the concepts underlying diagnostic research. We would also like to thank Professor Colin Begg, Professor Patrick Bossuyt, Mr Jon Deeks, Professor Constantine Gatsonis, Dr Khalid Khan, Dr Jeroen Lijmer, Dr David Moher, Professor Cynthia Mulrow, and Dr Gerben Ter Riet, for taking part in the Delphi procedure.

Authors' contributions

All authors contributed towards the conception and design of the project and read and approved the final manuscript. Penny Whiting (Research Fellow) worked as a reviewer on the project, carried out inclusion assessment, data extraction, synthesis of results and the Delphi procedure, as well as drafting the final report. Anne Rutjes (Research Fellow) worked as a reviewer on the project, carried out inclusion assessment, data extraction, synthesis of results and the Delphi procedure, as well as commenting on all drafts. Jacqueline Dinnes (Senior Research Fellow) worked as a reviewer on the project, carried out inclusion assessment, data extraction and synthesis of results, as well as commenting on all drafts. Patrick Bossuyt (Professor of Clinical Epidemiology) provided supervision for the project, commented on all drafts and obtained the funding for the project. Johannes Reitsma (Senior Clinical Epidemiologist) worked as a reviewer on the project, carried out inclusion assessment, data extraction and synthesis of results, as well as commenting on all drafts and providing supervision for the project. Jos Kleijnen (Director of Centre for Reviews and Dissemination) commented on all drafts, provided supervision for the project and obtained the funding for the project.

References

1. Deeks J. Systematic reviews of evaluations of diagnostic and screening tests. In Egger M, Davey Smith G, Altman D, editors. *Systematic reviews in health care: meta-analysis in context*. 2nd ed. London: BMJ Publishing; 2001. pp. 248–82.
2. Vamvakas EC. Meta-analyses of studies of the diagnostic accuracy of laboratory tests. A review of the concepts and methods. *Arch Pathol Lab Med* 1998;**122**:675–86.
3. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Stat Med* 1998;**17**:1033–53.
4. Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. *Br J Radiol* 1987;**60**:1071–81.
5. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ* 1986; **134**:587–94.
6. Sackett DL, Haynes RB, Guyatt GH, Tugwell T. The selection of diagnostic tests. In *Clinical epidemiology. A basic science for clinical medicine*. 2nd ed. London: Little, Brown; 1991. pp. 51–68.
7. Somoza E, Mossman D. Comparing and optimising diagnostic tests: an information-theoretical approach. *Med Decision Making* 1992; **12**:179–88.
8. Plasencia CM, Alderman BW, Baron AE, Rolfs RT, Boyko EJ. A method to describe physician decision thresholds and its application in examining the diagnosis of coronary artery disease based on exercise treadmill testing. *Med Decision Making* 1992;**12**:204–12.
9. Moons KG. *Diagnostic research: theory and application*. Rotterdam: Erasmus University; 1996.
10. Moons KG, Stijnen T, Michel BC, Buller HR, Van Es GA, Grobbee DE, *et al*. Application of treatment thresholds to diagnostic-test evaluation: an alternative to the comparison of areas under receiver operating characteristic curves. *Med Decision Making* 1997;**17**:447–54.
11. Thornbury JR. Clinical efficacy of diagnostic imaging: love it or leave it. *AJR Am J Roentgenol* 1994;**162**:1–8.
12. Simon DG, Lubin MF. Cost-effectiveness of computerized tomography and magnetic resonance imaging in dementia. *Med Decision Making* 1985;**5**:335–54.
13. Hobbs FD, Delaney BC, Fitzmaurice DA, Wilson S, Hyde CJ, Thorpe GH, *et al*. A review of near patient testing in primary care. *Health Technol Assess* 1997;**1**(5).
14. Munro J, Booth A, Nicholl J. Routine preoperative testing: a systematic review of the evidence. *Health Technol Assess* 1997;**1**(12).
15. Murray J, Cuckle H, Taylor G, Hewison J. Screening for fragile X syndrome. *Health Technol Assess* 1997;**1**(4).
16. Selly S, Donovan J, Faulkner A, Coast J, Gillatt D. Diagnosis, management and screening of early localised prostate cancer. *Health Technol Assess* 1997;**1**(2).
17. Berry E. A systematic literature review of spiral and ultrafast CT. *Health Technol Assess* 1999;**3**(18).
18. Gulliford M. Monitoring blood glucose control in diabetes mellitus: a systematic review. *Health Technol Assess* 2000;**4**(12).
19. Berry E. Intravascular ultrasound-guided interventions in coronary artery disease: a systematic literature review, with decision-analytic modelling, of outcomes and cost-effectiveness. *Health Technol Assess* 2000;**4**(35).
20. Price C. Systematic review of the use of biochemical markers of myocardial injury. *Health Technol Assess* (forthcoming).
21. Kleijnen J. A systematic review of tests for the diagnosis and evaluation of urinary tract infection in children under five years. *Health Technol Assess* (forthcoming).
22. Roderick P, Davies R, Raftery J, Crabbe D, Pearce R, Bhandari P, *et al*. The cost-effectiveness of screening for *Helicobacter pylori* to reduce mortality and morbidity from gastric cancer and peptic ulcer disease: a discrete-event simulation model. *Health Technol Assess* 2003;**7**(6).
23. Newman D. Systematic review on urine albumin testing for early detection of diabetic complications. *Health Technol Assess* (forthcoming).
24. Egger M. A study to evaluate the most cost effective way to screen for *Chlamydia trachomatis* genital tract infection and reduce its prevalence and associated burden of disease. *Health Technol Assess* (forthcoming).
25. Wyatt J. Does algorithm-guided diagnosis improve the clinical management of acute abdominal pain? A systematic review. *Health Technol Assess* (forthcoming).

26. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995;**274**:645–51.

27. Sheps SB, Schechter MT. The assessment of diagnostic tests: a survey of current medical research. *JAMA* 1984;**252**:2418–22.

28. Arroll BA, Schechter MT, Sheps SB. The assessment of diagnostic tests: a comparison of medical literature in 1982 and 1985. *J Gen Intern Med* 1995;**3**:443–7.

29. Beam CA, Sostman HD, Zheng JY. Status of clinical MR evaluations 1985–1988: baseline and design for further assessments. *Radiology* 1991; **180**:265–70.

30. Cooper LS, Chalmers TC, McCally M, Berrier J, Sacks HS. The poor quality of early evaluations of magnetic resonance imaging. *JAMA* 1988; **259**:3277–80.

31. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, *et al*. Does the quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998; **352**:609–13.

32. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273**:408–12.

33. 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**:1061–6.

34. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;**48**:119–30.

35. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA* 1994;**271**:389–91.

36. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. *JAMA* 1994;**271**:389–91.

37. Zhou XH. Correcting for verification bias in studies of a diagnosic test's accuracy. *Stat Methods Med Res* 1998;**7**:337–53.

38. Hui SL, Zhou XH. Evaluation of diagnosic tests without gold standards. *Stat Methods Med Res* 1998;**7**:354–70.

39. Nierenberg AA, Feinstein AR. How to evaluate a diagnostic marker test. Lessons from the rise and fall of dexamethasone suppression test. *JAMA* 1988;**259**:1699–702.

40. Kobberling J, Trampisch HJ, Windeler J. Memorandum for the evaluation of diagnostic measures. *J Clin Chem Clin Biochem* 1990;**28**:873–9.

41. van der Schouw YT, Verbeek AL, Ruijs JH. ROC curves for the initial assessment of new diagnostic tests. *Fam Pract* 1992;**9**:506–11.

42. van der Schouw YT, Verbeek AL, Ruijs SH. Guidelines for the assessment of new diagnostic tests. *Invest Radiol* 1995;**30**:334–40.

43. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;**39**:561–77.

44. Begg CB. Experimental design of medical imaging trials. Issues and options. *Invest Radiol* 1989;**24**:934–6.

45. Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. *J Gen Intern Med* 1989;**4**:288–95.

46. Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests. Recommended methods. Updated 6 June 1996. URL:http://www.cochrane.org/cochrane/sadtdoc1.htm. Accessed 23 April 2004.

47. Anonymous. How to read clinical journals: II. To learn about a diagnostic test. *CMAJ* 1981;**124**:703–10.

48. Greenhalgh T. How to read a paper. Papers that report diagnostic or screening tests. *BMJ* 1997; **315**:540–3.

49. Lijmer JG, Bossuyt PM, Mol BW, Bonsel GJ, van der Meulen JH, Prins MH. Evaluating diagnostic technologies: a review of the strategies. In 2nd International Conference. Scientific Basis of Health Services and 5th Annual Cochrane Colloquium, 8–12 October 1997, Amsterdam, 1997.

50. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomised trials into meta-analysis. *J Clin Epidemiol* 1992;**45**:255–65.

51. Tritchler D. Modelling study quality in meta-analysis. *Stat Med* 1999;**18**:2135–45.

52. NHS Centre for Reviews and Dissemination. Undertaking systematic reviews of research on effectiveness. York: University of York; 1996. CRD Report No. 4. URL:http://www.york.ac.uk/inst/crd/report4.htm. Accessed 23 April 2004.

53. 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**:417–22.

54. Sackett DL, Haynes RB. The architecture of diagnostic research. In Knottnerus J, editor. *The evidence base of clinical diagnosis*. London: BMJ Books; 2002. pp. 19–38.

55. Egglin TKP, Feinstein AR. Context bias: a problem in diagnostic radiology. *JAMA* 1996;**276**:1752–5.

56. Weller SC, Mann NC. Assessing rater performance without a 'gold standard' using consensus theory [published erratum appears in *Med Decis Making* 1997;**17**:240]. *Med Decis Making* 1997;**17**:71–9.

57. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J Clin Epidemiol* 1988; **41**:923–37.

58. Begg CB, Metz CE. Consensus diagnoses and 'gold standards' [published erratum appears in *Med Decis Making* 1990;**10**:149] [comment]. *Med Decis Making* 1990;**10**:29–30.

59. Henkleman RM, Kay I, Bronskill J. Receiver operating characteristic (ROC) analysis without truth. *Med Decis Making* 1990;**10**:24–9.

60. Buck AA, Gart JJ. Comparison of a screening test and a reference test in epidemiologic studies. I. Indices of agreement and their relation to prevalence. *Am J Epidemiol* 1966;**83**:586–92.

61. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987;**6**:411–23.

62. 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**:1028–35.

63. Begg C, Greenes R. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983;**39**:207–15.

64. 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**:405–10.

65. Panzer R, Suchman A, Griner P. Workup bias in prediction research. *Med Decis Making* 1987; **7**:115–19.

66. Ransohoff D, Feinstein A. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926–9.

67. Mower WR. Evaluating bias and variability in diagnostic test reports. *Ann Emerg Med* 1999; **33**:85–91.

68. Eikman EA, Cameron JL, Colman M, Natarajan TK, Dugal P, Wagner HN, Jr. A test for patency of the cystic duct in acute cholecystitis. *Ann Intern Med* 1975;**82**:318–22.

69. Meadway J, Nicolaides AN, Walker CJ, O'Connell JD. Value of Doppler ultrasound in diagnosis of clinically suspected deep vein thrombosis. *BMJ* 1975;**4**:552–4.

70. Ennis JT, Walsh MJ, Mahon JM. Value of infarct-specific isotope (^{99m}Tc-labelled stannous pyrophosphate) in myocardial scanning. *BMJ* 1975;**3**:517–20.

71. Newcombe R. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* 1998;**17**:857–72.

72. Harrell FE, Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; **15**:361–87.

73. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Evidence base of clinical diagnosis: designing studies to ensure that estimates of test accuracy are transferable. *BMJ* 2002;**324**:669–71.

74. Ransohoff D, Muir W. Diagnostic work-up bias in the evaluation of a test: serum ferritin and hereditary hemochromatosis. *Med Decis Making* 1982;**2**:139–46.

75. Panzer RJ, Kido DK, Hindmarsh T. A methodologic assessment of studies comparing magnetic resonance imaging and computed tomography of the brain. *Acta Radiol Suppl* 1986; **369**:269–74.

76. 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**:837–43.

77. 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**:699–710.

78. Detrano R, Gianrossi R, Mulvihill D, Lehmann K, Dubach P, Colombo A, *et al*. Exercise-induced ST segment depression in the diagnosis of multivessel coronary disease: a meta analysis. *J Am Coll Cardio* 1989;**14**:1501–8.

79. 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**:1289–95.

80. 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**:64–71.

81. Levy D, Labib S, Anderson K. Determinant of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. *Circulation* 1990;**81**:815–20.

82. Melbye H, Straume B. The spectrum of patients strongly influences the usefulness of diagnostic

tests for pneumonia. *Scand J Prim Health Care* 1993;**11**:241–6.

83. 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**:12–17.

84. 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. p. 90.

85. 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**:741–7.

86. 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**:751–5.

87. 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**:1461–7.

88. 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**:353–62.

89. 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**:2467–70.

90. van Rijkom HM, Verdonschot EH. Factors involved in validity measurements of diagnostic tests for approximal caries – a meta-analysis. *Caries Res* 1995;**29**:364–70.

91. Doubilet P, Herman P. Interpretation of radiographs: effect of clinical history. *AJR Am J Roentgenol* 1981;**137**:1055–8.

92. 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**:135–40.

93. 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**:140–4.

94. 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**:518–22.

95. 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.

96. Arana GW, Zarzar MN, Baker E. The effect of diagnostic methodology on the sensitivity of the TRH stimulation test for depression: a literature review. *Biol Psychiatry* 1990;**28**:733–7.

97. 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**:476–81.

98. 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.

99. Phelps CE, Hutson A. Estimating diagnostic test accuracy using a 'fuzzy gold standard'. *Med Decis Making* 1995;**15**:44–57.

100. Thibodeau L. Evaluating diagnostic tests. *Biometrics* 1981;**37**:801–4.

101. 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**:965–74.

102. 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**:18–24.

103. 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**:735–42.

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

105. Diamond GA. Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem. *Med Decis Making* 1992;**12**:22–31.

106. 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**:391–8.

107. 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. p. 31.

108. 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**:864–9.

109. Zhou XH. Effect of verification bias on positive and negative predictive values. *Stat Med* 1994; **13**:1737–45.

110. 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**:507–11.

111. 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**:85–9.

112. Elmore J, Wells C, Howard D. The impact of clinical history on mammographic interpretations. *JAMA* 1997;**277**:49–52.

113. Good B, Cooperstein L, DeMarino G. Does knowledge of the clinical history affect the accuracy of chest radiograph interpretation? *AJR Am J Roentgenol* 1990;**154**:709–12.

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, 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**:78–83.

116. Schreiber M. The clinical history as a factor in roentgenogram interpretation. *JAMA* 1963; **185**:137–9.

117. Berbaum KS, Franken EA, Jr, el-Khoury GY. Impact of clinical history on radiographic detection of fractures: a comparison of radiologists and orthopedists. *AJR Am J Roentgenol* 1989; **153**:1221–4.

118. Ciccone G, Vineis P, Frigerio A, Segnan N. Inter-observer and intra-observer variability of mammogram interpretation: a field study. *Eur J Cancer* 1992;**28A**:1054–8.

119. 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**:518–20.

120. 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**:458–65.

121. Cuaron A, Acero AP, Cardenas M, Huerta D, Rodriguez A, de Garay R. Interobserver variability in the interpretation of myocardial images with Tc-99m-labeled diphosphonate and pyrophosphate. *J Nucl Med* 1980;**21**:1–9.

122. Elmore J, Wells C, Lee C. Variability in radiologists' interpretation of mammograms. *N Engl J Med* 1994;**331**:1493–9.

123. 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**:588–93.

124. 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**:151–8.

125. Loy CT, Irwig LM, Katelaris PH, Talley NJ. Do commercial serological kits for *Helicobacter pylori* infection differ in accuracy? A meta-analysis. *Am J Gastroenterol* 1996;**91**:1138–44.

126. Badgett RG, Lucey CR, Mulrow CD. Can the clinical examination diagnose left-sided heart failure in adults? *JAMA* 1997;**277**:1712–19.

127. Holleman D, Simel D. Does the clinical examination predict airflow limitations? *JAMA* 1995;**273**:313–19.

128. Huicho L, Campos M, Rivera J, Guerrant RL. Fecal screening tests in the approach to acute infectious diarrhea: a scientific overview. *Pediatr Infect Dis J* 1996;**15**:486–94.

129. Heffner JE, Brown LK, Barbieri C, Deleo JM. Pleural fluid chemical-analysis in parapneumonic effusions: a metaanalysis. *Am J Respir Criti Care Med* 1995;**151**:1700–8.

130. Irwig L, Tostesen ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, *et al*. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;**120**:667–76.

131. Kent DL, Larson EB. Disease, level of impact, and quality of research methods. Three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. *Invest Radiol* 1992;**27**:245–54.

132. Wilson J, Junger G. *Principles and practice of screening for disease*. Geneva: World Health Organisation; 1968.

133. Hoffman RM, Kent DL, Deyo RA. Diagnostic accuracy and clinical utility of thermography for lumbar radiculopathy. A meta-analysis. *Spine* 1991; **16**:623–8.

134. Fahey MT, Irwig L, Macaskill P. Meta-analysis of Pap test accuracy. *Am J Epidemiol* 1995;**141**:680–9.

135. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. *Radiology* 1988;**167**:565–9.

136. Lacasse Y, Wong E, Guyatt GH, Cook DJ. Transthoracic needle aspiration biopsy for the diagnosis of localised pulmonary lesions: a meta-analysis. *Thorax* 1999;**54**:884–93.

137. Chien PFW, Khan KS, Ogston S, Owen P. The diagnostic accuracy of cervico-vaginal fetal fibronectin in predicting preterm delivery: an overview. *Br J Obstet Gynaecol* 1997;**104**:436–44.

138. Deyo R, Haselkorn J, Hoffman RM, Kent D. Designing studies of diagnostic tests for low back pain or radiculopathy. *Spine* 1994; **19**(Suppl):S2057–65.

139. van Tulder MW, Assendelft WJ, Koes BW, Bouter LM. Spinal radiographic findings and nonspecific low back pain: a systematic review of observational studies. *Spine* 1997;**22**:427–34.

140. Dunn G, Everitt B. *Clinical biostatistics: an introduction to evidence-based medicine*. London: Edward Arnold; 1995.

141. Feinstein AR. *Clinical epidemiology: the architecture of clinical research*. Philadelphia, PA: WB Saunders; 1985.

142. Fiellin D, Reid M, O'Connor P. Screening for alcohol problems in primary care. *Arch Intern Med* 2000;**160**:1977–89.

143. Guyatt G, Oxman A, Ali M, Willan A, McIlroy W, Patterson C. Laboratory diagnosis of iron-deficiency anaemia: an overview. *J Gen Intern Med* 1992;**7**:145–53.

144. van den Hoogen HM, Koes BW, van Eijk JT, Bouter LM. On the accuracy of history, physical examination, and erythrocyte sedimentation rate in diagnosing low back pain in general practice. *Spine* 1995;**20**:318–27.

145. Owens DK, Holodniy M, Garber AM, Scott J, Sonnad S, Moses L, *et al*. Polymerase chain reaction for the diagnosis of HIV infection in adults. A meta-analysis with recommendations for clinical practice and study design. *Ann Intern Med* 1996;**124**:803–15.

146. Rao JK, Weinberger M, Oddone EZ, Allen NB, Landsman P, Feussner JR. The role of antineutrophil cytoplasmic antibody (c-ANCA) testing in the diagnosis of Wegener granulomatosis. A literature review and meta-analysis. *Ann Intern Med* 1995;**123**:925–32.

147. Heffner JE, Brown LK, Barbieri CA. Diagnostic value of tests that discriminate between exudative and transudative pleural effusions. *Chest* 1997; **111**:970–80.

148. Koumans EH, Johnson RE, Knapp JS, St Louis ME. Laboratory testing for *Neisseria gonorrhoeae* by recently introduced nonculture tests: a performance review with clinical and public health considerations. *Clin Infect Dis* 1998;**27**:1171–80.

149. Nuovo J, Melnikow J, Hutchison B, Paliescheskey M. Is cervicography a useful diagnostic test? A systematic overview of the literature. *J Am Board Fam Pract* 1997;**10**:390–7.

150. Flynn K, Adams E, Anerson D. *Positron emission tomography: systematic review*. Veterans Affairs Medical Center Health Services Research & Development Service, Management Decision & Research Center. 1996; Report No. A6.

151. Kent DL, Haynor DR, Larson EB, Deyo RA. Diagnosis of lumbar spinal stenosis in adults: a metaanalysis of the accuracy of CT, MR, and myelography. *AJR Am J Roentgenol* 1992;**158**: 1135–44.

152. Kent DL, Haynor DR, Longstreth W, Larson E. The clinical efficacy of magnetic resonance imagining in neuroimaging. *Ann Intern Med* 1994; **120**:856–71.

153. Meade M, Richardson W. Selecting and appraising studies for a systematic review. *Ann Intern Med* 1997;**127**:531–7.

154. Owens DK, Holodniy M, McDonald TW, Scott J, Sonnad S. A meta-analytic evaluation of the polymerase chain reaction for the diagnosis of HIV infection in infants. *JAMA* 1996;**275**:1342–8.

155. Philbrick J, Horowitz R, Feinstein A. Methodologic problems of exercise testing for coronary artery disease: groups, analysis and bias. *Am J Cardiol* 1980;**46**:807–12.

156. Mullins M, Becker D, Hagspiel K, Philbrick J. The role of spiral volumetric computed tomography in the diagnosis of pulmonary embolism. *Arch Intern Med* 2000;**160**:293–8.

157. Devous MD, Sr, Thisted RA, Morgan GF, Leroy RF, Rowe CC. SPECT brain imaging in epilepsy: a meta-analysis. *J Nucl Med* 1998;**39**:285–93.

158. Hallan S, Asberg A. The accuracy of C-reactive protein in diagnosing acute appendicitis. *Scand J Clin Lab Invest* 1997;**57**:373–80.

159. Rappeport ED, Mehta S, Wieslander SB, Schwarz Lausten G, Thomsen HS. MR imaging before arthroscopy in knee joint disorders? *Acta Radiol* 1996;**37**:602–9.

160. Anand SS, Wells PS, Hunt D, Brill-Edwards P, Cook D, Ginsberg JS. Does this patient have deep vein thrombosis? *JAMA* 1998;**279**:1094–9.

161. Attia J, Margetts P, Guyatt G. Diagnosis of thyroid disease in hospitalized patients: a systematic review. *Arch Intern Med* 1999;**159**:658–65.

162. Bachmann MO, Nelson SJ. Impact of diabetic retinopathy screening on a British district population: case detection and blindness prevention in an evidence-based model. *J Epidemiol Community Health* 1998;**52**:45–52.

163. Badgett RG, Mulrow CD, Otto PM, Ramirez G. How well can the chest radiograph diagnose left ventricular dysfunction? *J Gen Intern Med* 1996; **11**:625–34.

164. Barlow J, Stewart-Brown S, Fletcher J. Systematic review of the school entry medical examination. *Arch Dis Child* 1998;**78**:301–11.

165. Bastian LA, Piscitelli JT. Is this patient pregnant? Can you reliably rule in or rule out early pregnancy by clinical examination? *JAMA* 1997;**278**:586–91.

166. Bastian LA, Nanda K, Hasselblad V, Simel DL. Diagnostic efficiency of home pregnancy test kits: a meta-analysis. *Arch Fam Med* 1998;**7**:465–9.

167. Becker D, Philbrick J, Bachhuber T, Humphries J. D-Dimer testing and acute venous thromboembolism. *Arch Intern Med* 1996;**156**:939–46.

168. Bell R, Petticrew M, Luengo S, Sheldon TA. Screening for ovarian cancer: a systematic review. *Health Technol Assess* 1998;**2**(2):1–84.

169. Bonis PA, Ioannidis JP, Cappelleri JC, Kaplan MM, Lau J. Correlation of biochemical response to interferon alfa with histological improvement in hepatitis C: a meta-analysis of diagnostic test characteristics. *Hepatology* 1997;**26**:1035–44.

170. Bradley KA, Boyd-Wickizer J, Powell SH, Burman ML. Alcohol screening questionnaires in women: a critical review. *JAMA* 1998;**280**:166–71.

171. Buntinx F, Wauters H. The diagnostic value of macroscopic haematuria in diagnosing urological cancers: a meta-analysis. *Fam Pract* 1997;**14**:63–8.

172. Conde-Agudelo A, Kafury-Goeta AC. Triple-marker test as screening for Down syndrome: a meta-analysis. *Obstet Gynecol Surv* 1998;**53**:369–76.

173. Da Silva O, Ohlsson A, Kenyon C. Accuracy of leukocyte indices and C-reactive protein for diagnosis of neonatal sepsis: a critical review. *Pediatr Infect Dis J* 1995;**14**:362–6.

174. De Bernardinis M, Violi V, Roncoroni L, Boselli AS, Giunta A, Peracchia A. Discriminant power and information content of Ranson's prognostic signs in acute pancreatitis: a meta-analytic study. *Crit Care Med* 1999;**27**:2272–83.

175. de Vries SO, Hunink MG, Polak JF. Summary receiver operating characteristic curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. *Acad Radiol* 1996;**3**: 361–9.

176. Hrung J, Sonad S, Schwartz J, Langlotz C. Accuracy of MR imagining in the work-up of suspicious breast lesions: a diagnostic meta-analysis. *Acad Radiol* 1999;**6**:387–97.

177. Kearon C, Julian JA, Newman TE, Ginsberg JS. Noninvasive diagnosis of deep venous thrombosis. *Ann Intern Med* 1998;**128**:663–77.

178. Koelemay MJ, Denhartog D, Prins MH, Kromhout JG, Legemate DA, Jacobs MJ. Diagnosis of arterial disease of the lower extremities with duplex ultrasonography. *Br J Surg* 1996;**83**:404–9.

179. Lederle FA, Simel DL. Does this patient have abdominal aortic aneurysm? *JAMA* 1999; **281**:77–82.

180. Liedberg J, Panmekiate S, Petersson A, Rohlin M. Evidence-based evaluation of three imaging methods for the temporomandibular disc. *Dentomaxillofacial Radiology* 1996;**25**:234–41.

181. Littenberg B, Siegel A, Tosteson ANA, Mead T. Clinical efficacy of SPECT bone imaging for low back pain. *J Nucl Med* 1995;**36**:1707–13.

182. Mayer J. Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma. *Med J Aust* 1997;**167**:206–10.

183. McGee S, Abernethy WB, Simel DL. Is this patient hypovolemic? *JAMA* 1999;**281**:1022–9.

184. Metlay JP, Kapoor WN, Fine MJ. Does this patient have community-acquired pneumonia? Diagnosing pneumonia by history and physical examination. *JAMA* 1997;**278**:1440–5.

185. Mol BW, Dijkman B, Wertheim P, Lijmer J, van der Veen F, Bossuyt PM. The accuracy of serum chlamydial antibodies in the diagnosis of tubal pathology: a meta-analysis. *Fertility & Sterility* 1997;**67**:1031–7.

186. Mol BW, Lijmer JG, Ankum WM, van der Veen F, Bossuyt PM. The accuracy of single serum progesterone measurement in the diagnosis of ectopic pregnancy: a meta-analysis. *Human Reproduction* 1998;**13**:3220–7.

187. Mol BW, Bayram N, Lijmer JG, Wiegerinck MA, Bongers MY, van der Veen F, *et al*. The performance of CA-125 measurement in the detection of endometriosis: a meta-analysis. *Fertil Steril* 1998;**70**:1101–8.

188. Pearl WS, Todd KH. Ultrasonography for the initial evaluation of blunt abdominal trauma: a review of prospective trials. *Ann Emerg Med* 1996; **27**:353–61.

189. Pollitt RJ, Green A, McCabe CJ, Booth A, Cooper NJ, Leonard JV, *et al*. Neonatal screening for inborn errors of metabolism: cost, yield and outcome. *Health Technol Assess* 1997;**1**(7):1–203.

190. Reed WW, Byrd GS, Gates RH, Jr, Howard RS, Weaver MJ. Sputum Gram's stain in community-

acquired pneumococcal pneumonia. A meta-analysis. *West J Med* 1996;**165**:197–204.

191. Selker HP, Zalenski RJ, Antman EM, Aufderheide TP, Bernard SA, Bonow RO, *et al*. An evaluation of technologies for identifying acute cardiac ischemia in the emergency department: a report from a national heart attack alert program working group. *Ann Emerg Med* 1997;**29**:13–87.
192. Spencer-Green G, Alter D, Welch HG. Test performance in systemic sclerosis: anti-centromere and Anti-Scl-70 antibodies [review]. *Am J Med* 1997;**103**:242–8.
193. Swart P, Mol BW, van der Veen F, van Beurden M, Redekop WK, Bossuyt PM. The accuracy of hysterosalpingography in the diagnosis of tubal pathology: a meta-analysis. *Fertil Steril* 1995; **64**:486–91.
194. Tugwell P, Dennis DT, Weinstein A, Wells G, Shea B, Nichol G, *et al*. Laboratory evaluation in the diagnosis of Lyme disease: clinical guideline, part 2. *Ann Intern Med* 1997;**127**:1109–23.
195. Wells PS, Lensing AW, Davidson BL, Prins MH, Hirsh J. Accuracy of ultrasound for the diagnosis of deep venous thrombosis in asymptomatic patients after orthopedic surgery. A meta-analysis. *Ann Intern Med* 1995;**122**:47–53.
196. Whited JD, Grichnik JM. Does this patient have a mole or a melanoma? *JAMA* 1998;**279**:696–701.
197. Moher D, Cook D, Jadad A. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technol Assess* 1999;**3**:12.
198. Deeks J, Dinnes J, Sowden A, D'Amico R, Petticrew M, Altman A. Evaluating non-randomised intervention studies. *Health Technol Assess*;**7**:27.
199. Heffner JE, Feinstein D, Barbieri C. Methodologic standards for diagnostic test research in pulmonary medicine. *Chest* 1998;**114**:877–85.
200. Lensing AW, Hirsh J. ^{125}I-fibrinogen leg scanning: reassessment of its role for the diagnosis of venous thrombosis in post-operative patients. *Thromb Haemost* 1993;**69**:2–7.
201. Radack K, Park S. Is there a valid association between skin tags and colonic polyps: insights from a quantitative and methodologic analysis of the literature. *J Gen Intern Med* 1993;**8**:413–21.
202. Rothwell PM, Pendlebury ST, Wardlaw J, Warlow CP. Critical appraisal of the design and reporting of studies of imaging and measurement of carotid stenosis. *Stroke* 2000;**31**:1444–50.
203. van der Wurff P, Hagmeijer RHM, Meyne W. Clinical tests of the sacroiliac joint: a systematic methodological review. Part 1: Reliability. *Man Ther* 2000;**5**:30–6.
204. Windeler J, Richter K, Kobberling J. Description and evaluation of diagnostic-tests in German-language medical journals. *Schweiz Med Wochenschr* 1988;**118**:1437–41.
205. Kelly S, Berry E, Roderick P, Harris KM, Cullingworth J, Gathercole L, *et al*. A checklist for the identification of bias in studies of the diagnostic performance of medical imaging modalities. (forthcoming).
206. Becker D, Philbrick J, Abbitt P. Real-time ultrasonography for the diagnosis of lower extremity deep venous thrombosis: the wave of the future? *Arch Intern Med* 1989;**149**:1731–4.
207. Liddle J, Williamson M, Irwig L. Method for evaluating research and guideline evidence. Report No. 0943126444. Sydney: NSW Health Department; 1996.
208. Arrive L, Renard R, Carrat F, Belkacem A, Dahan H, Le Hir P, *et al*. A scale of methodological quality for clinical studies of radiologic examinations. *Radiology* 2000;**217**:69–74.
209. Khan KS, Dinnes J, Kleijnen J. Systematic reviews to evaluate diagnostic tests [review]. *Eur J Obstet Gynecol Reprod Biol* 2001;**95**:6–11.
210. Black WC. How to evaluate the radiology literature. *AJR Am J Roentgenol* 1990;**154**:17–22.
211. Jensen K, Abel U. Methodik diagnostischer Validierungsstudien. Fehler in der Studienplanung und Auswertung. *Med Klin* 1999;**94**:522–9.
212. Greiner M, Gardner IA. Epidemiologic issues in the validation of veterinary diagnostic tests. *Prev Vet Med* 2000;**45**:3–22.
213. Heffner JE. Evaluating diagnostic tests in the pleural space: differentiating transudates from exudates as a model. *Clin Chest Med* 1998;**19**:277–94.
214. Riegelman RK, Hirsch RP. *Studying a study and testing a test: how to read the health science literature*. 3rd ed. Boston, MA: Little Brown; 1996.
215. Sackett DL. *Evidence-based medicine: how to practise and teach EBM*. 2nd ed. Edinburgh: Churchill Livingstone; 2000.
216. Deeks JJ. Using evaluations of diagnostic tests: understanding their limitations and making the most of available evidence. *Ann Oncol* 1999; **10**:761–8.
217. Mant D. Testing a test: three critical steps. In Jones R, Kinmouth A, editors. *Critical reading for primary care*. Oxford: Oxford University Press; 1995. pp. 183–202.
218. Kraemer HC. *Evaluating medical tests: objective and quantitative guidelines*. Newbury Park: Sage; 1992.
219. Gifford DR, Cummings JL. Evaluating dementia screening tests: methodologic standards to rate their performance. *Neurology* 1999;**52:**224–7.

220. Thornbury JR, Kido DK, Mushlin AI, Phelps CE, Mooney C, Fryback DG. Increasing the scientific quality of clinical efficacy studies of magnetic-resonance-imaging. *Invest Radiol* 1991;**26**:829–35.

221. Sox H, Stern S, Owens D, Abrahms H. *Monograph of the Council on Health Care Technology, Institute of Medicine: Assessment of diagnostic technology in health care*. Washington, DC: National Academy Press; 1989.

222. Bruns DE, Huth EJ, Magid E, Young DS. Toward a checklist for reporting of studies of diagnostic accuracy of medical tests. *Clin Chem* 2000;**46**:893–95.

223. Lang TA, Secic M. *How to report statistics in medicine: annotated guidelines for authors, editors, and reviewers*. Philadelphia, PA: American College of Physicians; 1997.

224. Haynes R, Sackett DL. Purpose and procedure (abbreviated). *Evidence Based Medicine* 1995;**1**:2.

225. Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;**323**:157–62.

226. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press; 1995.

227. Jadad AR, Moore A, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, *et al*. Assessing the quality of reports of randomised clinical trials: is blinding necessary? *Control Clin Trials* 1996; **17**:1–12.

228. Assendelft JJ, Koes BW, van Tulder MW, Bouter LM. Scoring the quality of clinical trials [letter]. *JAMA* 2000;**283**:1421.

229. ter Riet G, Leffers P, Zeegers P. Scoring the quality of clinical trials [letter]. *JAMA* 2000;**283**:1421.

230. van den Broucke JP. Scoring the quality of clinical trials [letter]. *JAMA* 2000;**283**:1422.

231. Juni P, Egger M. Scoring the quality of clinical trials [letter]. *JAMA* 2000;**283**:1422–3.

232. Klassen T. Bias against quality scores. 2001. URL:http://bmj.com/cgi/eletters/323/7303/42. Accessed 20 June 2002.

233. Juni P. Authors' response: Why scales are unhelpful for assessing quality of trials. *BMJ*, 2001. URL:http://bmj.com/cgi/eletters/323/7303/42. Accessed 20 June 2002.

234. Juni P, Altman DG, Egger M. Assessing the quality of controlled trials. *BMJ* 2001;**323**:42–6.

235. Juni P, Witschi A, Bloch RM, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;**282**:1054–60.

236. Greenland S. Invited Commentary: A critical look at some popular meta-analytic methods. *Am J Epidemiol* 1994;**140**:290–6.

237. Feinstein AR. *Clinimetrics*. Newhaven, CT: Yale University Press; 1987.

238. Kerr M. *The Delphi Process*. Remote and Rural Areas Research Initiative, NHS in Scotland, 2001. URL:http://www.rararibids.org.uk/documents/bid79-delphi.htm. Accessed 12 November 2002.

239. The Delphi technique in pain research. Scottish Network for Chronic Pain Research, 2001. URL:http://www.sncpr.org.uk/delphi.htm. Accessed 12 November 2002.

240. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979; **2**:420–8.

241. Bossuyt P, Reitsma J, Bruns D, Gatsonis C, Glasziou P, Irwig L, *et al*. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003; **49**:7–18.

242. Bachmann M, Nelson S. *Screening for diabetic retinopathy: a quantitative overview of the evidence, applied to the populations of health authorities and boards*. University of Bristol: Health Care Evaluation Unit; 1996;46.

243. Guyatt GH. Critical-evaluation of radiologic technologies. *Journal of the Canadian Association of Radiologists – Journal de l'Association Canadienne des Radiologistes* 1992;**43**:6–7.

244. Hoffman PAM, Nelemans P, Kemerink GJ, Wilmink JT. Value of radiological diagnosis of skull fracture in the management of mild head injury: meta-analysis. *J Neurol Neurosurg Psychiatry* 2000; **68**:416–22.

245. Bruns DE, Huth EJ, Magid E, Young DS. Towards a checklist for reporting of studies of diagnostic accuracy. *Clin Chem* 1997;**43**:2211.

246. Greenes R, Begg C. Assessment of diagnostic technologies: methodology for unbiased estimation from samples of selectively verified patients. *Invest Radiol* 1985;**20**:751–6.

247. Shlipak MG, Lyons WL, Go AS, Chou TM, Evans GT, Browner WS. Should the electrocardiogram be used to guide therapy for patients with left bundle-branch block and suspected myocardial infarction? *JAMA* 1999;**281**:714–19.

248. Fineberg HV, Hiatt HH. Evaluation of medical practices. The case for technology assessment. *N Engl J Med* 1979;**301**:1086–91.

249. Sackett D. A primer on the precision and accuracy of the clinical examination. *JAMA* 1992; **267**:2638–44.

Appendix 1

Search strategies

MEDLINE

Searched 22 March 2001 via SilverPlatter
Date coverage: 1966 to December 2000
Search strategy
explode "Sensitivity-and-Specificity"/ all subheadings
explode "Mass-Screening"/ all subheadings
"Reference-Values"
"False-Positive-Reactions"
"False-Negative-Reactions"
specificit*
false negative
false positive
accuracy
screening
predictive value*
reference value*
likelihood ratio*
sroc
receiver operat* curve*
receiver operat* characteristic*
roc* in ti,ab,mesh
#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17
bias
#18 and #19

EMBASE

Searched 11 July 2001 via SilverPlatter
Date coverage: 1980 to June 2001
Search strategy
"sensitivity-and-sensibility"/ all subheadings
explode "mass-screening"/ all subheadings
"reference-value"/ all subheadings
explode "diagnostic-test"/ all subheadings
explode "laboratory-diagnosis"/ all subheadings
false positive reaction*
false negative reaction*
specificit*
false negative
false positive
screening
accuracy
predictive value*
reference value*
likelihood ratio*
sroc
receiver operat* characteristic*
roc* in ti,ab,de
"receiver-operating-characteristic"/ all subheadings
#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18 or #19
bias
#20 and #21
exact{EDITORIAL} in DT
exact{ERRATUM} in DT
exact{LETTER} in DT
exact{NOTE} in DT
#23 or #24 or #25 or #26
#22 not #27

BIOSIS Previews

Searched 10 September 2001 via EDINA, web service
Date coverage: 1985 to 7 September 2001
Search strategy
sensitivity n specificity
mass n screen*
reference n value*
(false n positive n reaction*)
(false n negative n reaction*)
specificit*
false n negative
false n positive
accuracy
screening
predictive n value*
reference n value*
likelihood n ratio*
sroc
(receiver n operat* n curve*)
(receiver n operat* n characteristic*)
roc*
1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17
bias
18 and 19

Cochrane Methodology Register

Searched 11 July 2001 via Cochrane Library 2001 (Issue 1)
Searched for papers that referred to "diagnos*".

DARE administrative database (an internal CRD database used in the production of DARE)

Searched 11 July 2001.
Searched for methodology papers that referred to "diagnos*".

Appendix 2

Data extraction tables: objective 1

Study details	Methods	Bias	Evidence provided[a]
Arana, 1990[96] **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 was compared between studies that used the DSM-III and the RDC as the reference standard	Absent or 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%)
Berbaum, 1989[110,117] **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 effect of knowledge of localising symptoms and signs in the detection of fractures was studied. 40 radiographs of the extremities were examined twice by 7 orthopaedic surgeons and 7 radiologists; the sessions were separated by 4–6 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 whether 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 ROC 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 'overread' or 'underread'). For orthopaedic surgeons the analysis of ROC parameters also found that clues regarding the location of trauma facilitate detection of fractures. The area under the ROC curve showed an 11% improvement ($p < 0.001$)
		Observer/ instrument variation	Statistical comparison of the two experiments showed that orthopaedic surgeons depend on clinical history much more than do radiologists. This was demonstrated by a statistically significant prompting speciality interaction ($p < 0.05$)
Bowler, 1998[102] **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. 192 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 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

continued

Study details	Methods	Bias	Evidence provided[a]
Boyko, 1988[97] **Study design**: numerical; 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 formulae 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	Absent or 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%*
Cecil, 1996[103] **Study design**: real life; diagnostic accuracy study, retrospective **Objective**: to determine the sensitivity, specificity, PPV and NPVs 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 database, reports of 4354 stress SPECT thallium studies from 1 January 1986 to 31 December 1992 were reviewed. All patients with a known history of MI 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%.	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[118] **Study design**: real life: experimental **Objective**: to evaluate the performance of radiologists in mammographic mass screening **Type of analysis**: statistical	7 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/ instrument 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 (interobserver comparisons). For the intraobserver 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 to 85.7 at the first reading and from 44.4 to 88.9 at the second reading. Specificity ranged from 52.9 to 73.5 at the first reading and from 50.0 to 80.0 at the second reading

continued

Study details	Methods	Bias	Evidence provided[a]
Cohen, 1987[119] **Study design**: real life; experimental **Objective**: to assess the influence of training and experience on the interpretation of FNAB specimens **Type of analysis**: statistical	50 cases were selected from the cytology registry of the University of California, San Francisco. Each case had histological 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, whereas 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 others were non-experts. ROC curves were used to investigate observer variability	Observer/ instrument variation	The ROC curves showed that training and experience significantly influenced interpretation of breast FNAB specimens. The 2 experts operated at a higher level of sensitivity and specificity than the 3 non-experts. Pairwise comparison of areas under the ROC curves showed significant differences between the experts and non-experts
Corley, 1997[120] **Study design**: real life; experimental **Objective**: to establish a histological diagnosis of pneumonia by consensus of a panel of pathologists, to test the interobserver and intraobserver variation in the histological diagnosis of pneumonia, to compare the diagnostic accuracy of diagnosing pneumonia with and without preselected histological criteria, and to establish more specific histological 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 after the initial evaluation, the same slides were resubmitted to one of the pathologists for re-evaluation to look for intraobserver error. Finally, the slides were reviewed and categorised by the criteria of Johanson and colleagues into no pneumonia, mild, moderate or severe bronchopneumonia. A comparison was made of the patients selected as demonstrating histological pneumonia by each of the examinations	Observer/ instrument 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 histological criteria of Johanson and colleagues, 14 patients were selected as having pneumonia compared with only nine patients selected by consensus of 3 of 4 pathologists. Unanimous decisions among the observers were present in 30 patients (77%)

continued

Study details	Methods	Bias	Evidence provided[a]
Cuaron, 1980[121] **Study design**: real life; experimental **Objective**: to determine the possible bias of experience on the correct interpretation of ^{99m}Tc-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 ^{99m}Tc-labelled 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, i.e. one with 5 criterion levels. Results were expressed as ROC curves, from which the optimal individual criterion level for each observer was determined	Observer/ instrument 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 programmes to evaluate periodically the performance of every physician who interprets studies, to establish his or her 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 and 90% between observers and specificity varied between 79 and 93%
Curtin, 1997[76] **Study design**: real life; diagnostic accuracy study, retrospective **Objective**: to evaluate the accuracy of 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 gender and weight subgroups	Variation by clinical and demographic subgroups	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 <60 kg and increased to 54.6% in the subgroup weighing >80 kg. The major increase in sensitivity for both genders occurred for participants weighing ≥ 80 kg. In the subgroup weighing >60 kg the sensitivity was higher in females than in males. In both genders and in all subgroups the specificity was 100%, but the lower bound of the 95% CI 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

continued

Study details	Methods	Bias	Evidence provided[a]
De Neef, 1987[98] **Study design**: numerical; 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	Absent/ 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 could 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[77] **Study design**: real life; review **Objective**: to investigate technical and methodological factors that may affect the reported accuracies of diagnostic tests of exercise thallium scintigraphy **Type of analysis**: statistical	To assess the influence of various factors on the accuracy of exercise thallium scintigraphy, the medical literature (1977–1986) was non-selectively searched and meta-analysis was applied to the 56 publications retrieved. These were analysed for year of publication, gender and mean age of patients, percentage of patients with angina pectoris, percentage of patients with prior MI, percentage of patients taking β-blocking medications, and for angiographic referral (work-up) bias, blinding of tests and technical factors	Variation by clinical and demographic subgroups	The percentage of men in the study group was independently significantly ($p < 0.05$) related to test sensitivity. Mean age, use of β-blocking medications and adequate definition of the study group did not significantly affect test sensitivity or specificity. The percentage of patients with MI had the highest correlation with sensitivity (0.45, $p = 0.0007$). The inclusion of patients with prior infarction was independently significantly ($p < 0.05$) related to test sensitivity. The mean sensitivity of studies that included prior infarctions was 86% compared with 79% in studies that excluded infarctions
		Distorted selection of participants	Avoidance of a limited challenge group did not significantly affect test sensitivity or specificity
		Partial verification bias	The presence of work-up bias adversely affected specificity ($p < 0.05$), i.e. specificity is higher in studies where work-up bias is present
		Review bias	Blinding of both the thallium scintigram and the coronary angiogram tended to decrease the concordance between the two; this effect was statistically significant only for sensitivity

continued

Study details	Methods	Bias	Evidence provided[a]
Detrano, 1988[79] **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 ANOVA. 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: percentage of 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 work-up bias, blinding of test and reference standard, technical details	Variation by clinical and demographic subgroups	The percentage of patients with prior MI had the highest correlation with sensitivity; sensitivity was highest in studies that included previous MI. Mean age and use of β-blocking medication did not affect test performance. Gender was significantly associated with sensitivity, but not with specificity. Percentage of men and previous MI were significantly associated with sensitivity in the multivariate analysis
			Adequate definition of the study group had non-significant effects on sensitivity and specificity
		Distorted selection of participants	Avoidance of limited challenge group had non-significant effects on sensitivity and specificity
		Absent or inappropriate reference standard	Angiographic disease verification was not significantly related to test performance. Sensitivity and specificity were higher in studies that used tomographic imaging, but only sensitivity was significantly higher. Tomographic imaging was significantly associated with sensitivity and specificity in the multivariate analysis
		Change in technology of test	Automation of the reading of the scintigraph improved sensitivity but decreased specificity; differences were significant
		Disease progression bias	The maximum interval between scintigraphy and angiography was not associated with test performance
		Difference in test protocol	Exercise protocol was not significantly related to test performance
		Partial verification bias	Work-up 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 with 86.6% in non-blinded studies

continued

Study details	Methods	Bias	Evidence provided[a]
Detrano, 1989[78] **Study design**: real life; review **Objective**: to evaluate the variability in the reported accuracy of the exercise 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	Variation by clinical and demographic subgroups	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 β-blocking agents
		Absent or 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 the following standard: avoidance of work-up bias was not associated with test performance
		Review bias	Whether the authors complied with all of the following standards: blind reading of angiogram, blind reading of exercise ECG, was not associated with test performance
		Inappropriate handling of uninterpretable/ indeterminate/ intermediate test 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

continued

Study details	Methods	Bias	Evidence provided[a]
Diamond, 1992[105] **Study design**: numerical; 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 percentage 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 that selection bias significantly distorts the determination of predictive accuracies calculated by Bayes' theorem, and that these distortions can be significantly offset by a correction algorithm*
Diamond, 1991[104] **Study design**: numerical; modelling **Objective**: to assess the ability of the Begg–Greenes[63] 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 was 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 an 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; and (3) the area under the ROC curve is insensitive 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*

continued

Study details	Methods	Bias	Evidence provided[a]
Doubilet, 1981[91] **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 workload of readers who were unaware that a study was being carried out. 8 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 4 readings with a suggestive and 4 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 compared with 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[55] **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	A test set of 24 right pulmonary arteriograms with a 33% prevalence of 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%. 6 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 ROC curves were compared with the paired *t*-test	Disease prevalence/ severity	In the context of group A's higher disease prevalence, radiologists shifted more of their diagnoses toward higher suspicion than expected by chance ($p = 0.03$, sign test). In group A, mean sensitivity for diagnosing PE was significantly higher (75% vs 60%; $p = 0.04$), and the area under the ROC curve was significantly larger (0.88 vs 0.82; $p = 0.02$). 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 in obtaining unbiased estimates of diagnostic accuracy for both new and existing technologies. Overall specificity was similar in both groups (64% vs 68%)
Eldevick, 1982[111] **Study design**: real life; experimental **Objective**: to assess the effect of clinical bias on the interpretation of myelography and spinal CT **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 out of 107 CT interpretations and 12 out of 103 myelographic interpretations differed between the first and second readings. 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 for observers to interpret questionable myelographic or CT 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

continued

Study details	Methods	Bias	Evidence provided[a]
Elmore, 1994[122] **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. 10 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/ instrument 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 work-up ranged from 74 to 96% for mammograms from the women with cancer and from 11 to 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[112] **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 washout period, 10 radiologists read mammograms for the same 100 women, randomly divided into 2 groups of 50. For one 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 work-up: an alerting history (e.g. breast symptoms or family history of breast cancer) increased the number of work-ups recommended in patients without cancer ($p = 0.01$) and a non-alerting history led to fewer recommended work-ups 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 radiologist to change a previous biopsy recommendation
Froelicher, 1998[101] **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 work-up 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 were calculated and compared with other similar studies conducted in populations where work-up bias was present	Change in technology of the test	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

continued

Study details	Methods	Bias	Evidence provided[a]
Good, 1990[113] **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 4 board-certified radiologists, both with and without accompanying clinical histories, was 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 ROC curves showed that, with the exception of interpretation of one abnormality by one radiologist, there were no statistically significant differences ($p < 0.05$) 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
Hlatky, 1984[80] **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 were studied. 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	Variation by clinical and demographic subgroup	5 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 gender. Only maximal exercise heart rate had a significant, independent effect on exercise electrocardiographic specificity
Lachs, 1992[92] **Study design**: real life; diagnostic accuracy study, prospective **Objective**: to determine whether the leucocyte esterase and bacterial nitrite rapid dipstick test for UTI is susceptible to spectrum bias **Type of analysis**: statistical	366 consecutive adult patients in whom clinicians performed urinalysis to diagnose or exclude UTI were studied in an urban emergency department and walk-in clinic. After the patient encounter, but before a 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 leucocyte 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/ severity	In the 107 patients with a high (>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 (≤ 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 to 0.57) and 0.78 (0.73 to 0.79), respectively The leucocyte 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

continued

Study details	Methods	Bias	Evidence provided[a]
Levy, 1990[81] **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 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 genders in the sensitivity and specificity of the ECG for echocardiographically defined LVH. The Cochran–Mantel–Haenszel statistic was used to adjust for gender and test the association between cigarette smoking and sensitivity and specificity of the ECG. Bivariate logistic regression was used to adjust for gender and to test the sensitivity and specificity trends with increasing age, obesity and left ventricular mass/height	Variation by clinical and demographic subgroup	Influence of gender: sensitivity was marginally lower in women (5.6 vs 9%, $p = 0.075$); specificity was high in both genders (99.4% in women and 98.1% in men)
			Influence of age: there was a trend for sensitivity to increase with increasing age ($p < 0.0001$, gender adjusted); there was a trend for specificity to decline with advancing age ($p < 0.001$, gender adjusted)
			Influence of obesity: sensitivity was inversely related to increasing BMI ($p < 0.05$ for trend, gender adjusted); no specific differences in specificity were observed
			Influence of smoking: sensitivity was lower among smokers than non-smokers (5.7% vs 10.9% in women, 1.6% vs 8% in women; $p = 0.001$ gender-adjusted). There were no statistically significant differences in specificity
		Disease prevalence/ severity	Influence of severity of LVH: a statistically significant trend towards increasing sensitivity of the ECG with increasing severity of LVH was observed for both genders ($p < 0.001$)
Lijmer, 1999[33] **Study design**: real life; review **Objective**: to determine empirically 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. Associations between study characteristics and estimates of diagnostic accuracy were evaluated with a regression model. The RDOR, which compared the DORs 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	Variation by clinical and demographic subgroups	Diagnostic performance was overestimated when no description of the population under study was provided (RDOR 1.4, 95% CI 1.1 to 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 to 4.5)
			Non-consecutive patient enrolment did not have any significant effect on diagnostic performance (RDOR 0.9, 95% CI 0.7 to 1.1), neither did a retrospective study design (RDOR 1.0, 95% CI 0.7 to 1.4)
		Difference in test protocol	When no criteria for the test were described diagnostic performance was overestimated (RDOR 1.7, 95% CI 1.1 to 2.5)
			When no criteria for the reference standard execution were described diagnostic performance was underestimated (RDOR 0.7, 95% CI 0.6 to 0.9)

continued

Study details	Methods	Bias	Evidence provided[a]
		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 to 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 to 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 to 1.9)
Lijmer, 1996[106] **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 years 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 LR was closer to 1.0 than the LR calculated from the verified sample
Melbye, 1993[82] **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, CRP analysis and ESR, in the diagnosis of pneumonia **Type of analysis**: statistical	The diagnostic properties (sensitivity, specificity, LR and PPV) of the cues compared with radiographic pneumonia were evaluated for the following groups: (1) all 581 included patients; (2) 402 patients who also underwent physical chest examination; (3) 188 patients classified by the doctors as having a lower respiratory tract infection; (4) 79 patients referred for radiography by the doctors. Only 229 of patients had radiographs (reference standard) ordered by a 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	Variation by clinical and demographic subgroups	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 lower respiratory tract infection. 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

continued

Study details	Methods	Bias	Evidence provided[a]
Miller, 1998[107] **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 MI or revascularisation who underwent stress ^{201}Tl or ^{99m}Tc-sestamibi imaging were studied, 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 (work-up bias) led to an overestimation of sensitivity (estimated as 97%, 66% after mathematical correction) and an underestimation of specificity (estimated as 13%, corrected estimate 73%)
Mol, 1999[108] **Study design**: real life; review **Objective**: to evaluate the effect of verification bias on the accuracy of first trimester nuchal translucency measurement for Down's 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. 15 studies without and 10 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's 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
Moons, 1997[83] **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)	Variation by clinical and demographic subgroup	The sensitivity of the ST/HR depression substantially differed according to gender, expected workload, absolute achieved workload, and relative workload SBP at peak exercise. Variation with smoking, cholesterol level and baseline SBP was less marked. The specificity differed according to gender, 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 gender, smoking, cholesterol level, baseline SBP, relative workload and SBP at peak exercise
		Disease prevalence/ severity	The sensitivity of the ST-segment/heart rate depression varied according to number of disease vessels. Variation across patients with non-specific and atypical angina compared with typical angina was less marked

continued

Study details	Methods	Bias	Evidence provided[a]
Morise, 1994[84] **Study design**: real life; diagnostic accuracy study, prospective **Objective**: to investigate whether gender 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 using 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, gender, symptoms, diabetes, cholesterol and peak exercise heart rate. Positive exercise ST criterion was ≥ 1 mm horizontal/downsloping depression	Variation by clinical and demographic subgroups	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 with 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 work-up bias in men and women was equivalent
Morise, 1995[85] **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 gender-related differences in post-test referral bias, the accuracy of exercise electrocardiography was compared 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	Variation by clinical and demographic subgroups	Sensitivity and specificity were significantly greater in men than in women with use of the biased or unbiased groups. The amounts by which sensitivity decreased and specificity increased were 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 gender-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 for the biased group. These differences reflect the effects of post-test referral bias

continued

Study details	Methods	Bias	Evidence provided[a]
O'Connor, 1996[93] **Study design**: real life; diagnostic accuracy, prospective **Objective**: to investigate whether within the population of suspected MS patients, there would be differences in MRI and EP sensitivity and specificity between those with mild MS and 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. 204 patients also received EP testing. The group was divided into 'possible' and 'probable' MS subgroups; sensitivity and specificity for MRI and EP were calculated separately for these subgroups, and the differences between them investigated	Disease prevalence/ severity	The sensitivity of MRI in patients with suspected MS was 58% 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 ($p < 0.01$)
Panzer, 1987[65] **Study design**: real life; diagnostic accuracy study, prospective **Objective**: to explore the potential impact of work-up 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 neurological 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 work-up 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 1 of the 3 findings comprised the biased sample; the remaining 195 patients were excluded from the study population Sensitivity, specificity and LRs were calculated for various clinical predictors in both the biased and unbiased samples	Partial verification bias	The frequency of each of the 3 clinical predictors used to select the biased sample was increased in that sample; this led to increased sensitivity and decreased specificity in the biased 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 LRs for the findings used to select the sample were consistently smaller than the LRs in the unbiased sample; LRs for related findings were also decreased; results were inconsistent for unrelated findings

continued

Study details	Methods	Bias	Evidence provided[a]
Phelps, 1995[99] **Study design**: numerical; 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	Monte Carlo studies were used to define inaccurate diagnostic tests and inaccurate fuzzy reference standards by adding various amounts of random noise to the true reference standard results. ROC curves were then estimated using these synthetic 'diagnostic test' data and as the reference standard either the truth or the fuzzy reference standard results that measures the truth with error. The estimated ROC areas were compared to determine the consequences of having an imperfect reference standard and the possible gains from using methods to offset the inherent fuzzy gold standard bias	Absent or inappropriate reference standard	*The results show that: (1) when diagnostic test errors are statistically independent from inaccurate reference standard errors, estimated test accuracy declines; and (2) when the test and the fuzzy reference standard have statistically dependent errors, test accuracy can become overstated*
Philbrick, 1982[89] **Study design**: real life; diagnostic accuracy study, prospective **Objective**: 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	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 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. Reasons why some of the patients included in this study would not be included in a diagnostic evaluation study, and the theoretical effect that this would have on the estimates of test performance, are discussed	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
		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, 1 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

continued

Study details	Methods	Bias	Evidence provided[a]
		Inappropriate handling of uninterpretable/ indeterminate/ intermediate test results	If technically unsatisfactory exercise test results were excluded, 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 were removed from further consideration, 62% would be excluded
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 3 equal-size groups. Group I received cues directed to the correct diagnosis on 28 of 56 test posteroanterior 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 ($p < 0.05$)
Raab, 2000[115] **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 **Type of analysis**: statistical	97 bronchial brush specimens were selected retrospectively from cytology files. Each of the specimens consisted of 2 slides; all cases had histological 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 was provided, on the second circulation, 2–3 months later, clinical history was provided. Clinical history included comprised gender, age, clinical findings (if any) and clinical suspicion of disease. Each observer scored each case as definitely benign, probably benign, possibly 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 LR for the benign category was lower with clinical history than without clinical history, i.e. a benign diagnosis was more likely to indicate 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 effect. For example, for the malignant category, if clinical history was provided the LR increased for 2 observers and decreased for 3 observers. For each observer the PPV of a malignant diagnosis was similar if history was or was not provided. For each observer, the NPV 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 an ROC curve, of all pathologists increased if history was provided. For the pooled data across all pathologists there was a statistically significant difference ($p < 0.05$) between the accuracy of the diagnoses with history and without history

continued

Study details	Methods	Bias	Evidence provided[a]
Raab, 1995[123] **Study design**: real life; experimental **Objective**: to use the bronchial brush specimen as an example, to 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 2 slides; all cases had histological 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/ instrument 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[66] **Study design**: real life; review **Objective**: to determine why many diagnostic tests 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 CEA test in the diagnosis of colonic cancer and the 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 from 1969 to 1973 for articles on CEA and from 1968 to 1973 for articles on NBT. Papers that had no original data or fewer than 10 patients, and 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 prevalence/ severity	CEA: the 3 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 out of 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 for 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 not used in any of the 4 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 were interpreted relatively objectively NBT: this test is interpreted subjectively and has a high degree of observer variability. 3 studies contained precautions against biased test interpretation and only 2 tried to avoid biased diagnostic interpretation; only 1 of these studies found a high specificity for the test and none found a high sensitivity

continued

Study details	Methods	Bias	Evidence provided[a]
Ransohoff, 1982[74] **Study design**: real life; review of 2 studies **Objective**: to provide an empirical illustration of diagnostic work-up bias **Type of analysis**: narrative	Two major reports examined the utility of serum ferritin in detecting iron overload in relatives of patients with hereditary hemochromatosis. The reference standard is liver biopsy. Investigators from Brisbane found that ferritin was elevated in 15 out of 15 relatives with marked iron overload, as indicated by a histological 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 histological grading. This study aims to identify and assess possible reasons for these divergent results	Partial verification bias	In the Boston 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 >140 μg/100 ml. 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[64] **Study design**: real life; diagnostic accuracy study, prospective **Objective**: to determine the effects of gender and 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 echocardiograph 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	Variation by demographic characteristics	After correction for verification bias, sensitivity was lower in women than in 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, and specificity was 83% in men and 86% in women
Ronco, 1996[124] **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 identified through colpohistological and cytological 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 3 expert supervisors who reached a consensus diagnosis	Observer/ instrument variability	Sensitivity of the cytologists was less than that of the supervisors: they correctly diagnosed 30 out of 34 smears judged as positive by supervisors

continued

Study details	Methods	Bias	Evidence provided[a]
Rozanski, 1983[94] **Study design**: real life; diagnostic accuracy study, retrospective **Objective**: to verify the dramatic temporal decline in specificity of exercise radionuclide ventriculography and to determine its cause **Type of analysis**: statistical	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/ severity	Most patients studied in the early period had normal responses (94% for ejection fraction and 84% for wall motion). In contrast, normal responses were less frequent in patients studied in the recent period (49% for ejection fraction and 36% for wall motion; $p < 0.001$). The probability of coronary disease before testing was higher in these patients (38 vs 7%, $p < 0.001$). The temporal decline in specificity is partly a result of a change in the population being tested (pre-test referral bias)
		Partial verification bias	More patients studied in the recent period underwent radionuclide ventriculography before angiography (78 vs 22%, $p < 0.001$), and more of these prior studies had abnormal results than those performed after angiography (55 vs 6%, $p < 0.0001$). 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)
Santana-Boado, 1998[86] **Study design**: real life; diagnostic accuracy study, prospective **Objective**: to compare the diagnostic accuracy of SPECT between both genders and to assess the influence of analysing only the patients with coronary angiography instead of all the patients submitted to the study **Type of analysis**: statistical	702 consecutive patients without previous MI were studied with SPECT. 163 had coronary angiography (select minority) and 539 did not (silent majority). All patients underwent exercise stress testing and simultaneous dipyrimadole was administered in 32% of patients who did not achieve maximum predicted heart rates. Diagnostic accuracy of the test was calculated for the select minority, then sensitivity and specificity were recalculated according to the Diamond criteria	Variation by clinical and demographic subgroups	In verified patients sensitivity was lower in men than in women, but no gender difference in sensitivity was present after correction for verification bias
		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)

continued

Study details	Methods	Bias	Evidence provided[a]
Schreiber, 1963[116] **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, gender, 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 other 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 history, although this was not significant ($p = 0.18$)
Stein, 1993[87] **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 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 2 groups according to whether they had prior cardiac or pulmonary disease. Sensitivity, specificity and PPV of PE based on the cumulative number of mismatched segmental defects were calculated separately for patients with and without cardiopulmonary disease. These data were stratified according to whether patients underwent obligatory angiography or patient-requested angiography	Variation by clinical and demographic subgroups	At ≥ 0.5 mismatched segmental equivalents PPV was 80% among patients with no prior cardiopulmonary disease, compared with 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 without 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 2 groups. Areas under the ROC curve were higher for patients with no prior cardiopulmonary disease (0.8905 vs 0.8215)

continued

Study details	Methods	Bias	Evidence provided[a]
Steinbauer, 1998[88] **Study design**: real life; diagnostic accuracy study, prospective **Objective**: to test for ethnic and gender bias in 3 self-report screening tests for alcohol use disorders in a primary care population **Type of analysis**: statistical	Primary care patients were randomly selected from appointment lists of a university-based family practice clinic. A probability sample of 1333 adult family practice patients, stratified by gender and ethnicity, was studied. Patients completed a diagnostic interview to determine the presence of a current alcohol use disorder, and 3 screening tests: the CAGE questionnaire, the SAAST and the AUDIT	Variation by clinical and demographic subgroups	The areas under the 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 > 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 LRs for the AUDIT were similar to or higher than those for the other screening tests, whereas negative LRs 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 gender bias
Taube, 1990[95] **Study design**: numerical; 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; and (3) those with advanced disease in whom 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 using the example of a data set of 168 cases of epithelial ovarian cancer	Disease prevalence/ severity	Sensitivity calculated on all available data (i.e. for all 3 stages of disease combined) = 0.83; for the clearly malignant cases sensitivity = 0.96, for mucinous cases sensitivity = 0.46, and 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*
Thibodeau, 1981[100] **Study design**: numerical; 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 diseased. 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

continued

Study details	Methods	Bias	Evidence provided[a]
van der Schouw, 1995[53] **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 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 was selected	Disease prevalence/ severity	As the criteria used to select patients became stricter the test properties changed markedly. As the selection criteria were widened (and so disease prevalence decreased), both sensitivity and specificity increased; the LR+ increased significantly from 4 to 28
van Rijkom, 1995[90] **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 2 categories: strong and weak. *D* was calculated for each study. A multivariate ANOVA 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 originated from *in vivo* studies were higher than those from *in vitro* studies. In the multivariate analysis *D* values obtained from *in vivo* studies were significantly different to 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
		Absent or inappropriate reference standard	On average, weak validation methods yielded higher values of *D* than strong validation methods. In the multivariate analysis *D* values were not statistically significantly different between validation methods ($p > 0.05$)

continued

Study details	Methods	Bias	Evidence provided[a]
Zhou, 1994[109] **Study design**: numerical; modelling with example using diagnostic accuracy design **Objective**: to examine the effect of verification bias on PPVs and NPVs **Type of analysis**: statistical	The effect of verification bias on estimated PPVs and NPVs based on only patients with verified disease statuses (the so-called naive estimators) was studied. By applying the maximum likelihood method the magnitude of the biases of the naive estimators was 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 naive 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 naive estimators (using only verified cases) for the PPVs and NPVs are 88% (95% CI 84 to 92) and 67% (95% CI 57 to 77), respectively. The maximum likelihood estimators for the true range in PPVs and NPVs could range from 81 to 93% and from to 24 to 93%, respectively. For this example, the naive estimator for the PPVs is reasonably robust against violation of the conditional independence assumption, while the naive estimator of the NPV is sensitive to violation of the assumption

[a] Empirical evidence is reported in standard print, theoretical evidence is reported in italics.
AUDIT, Alcohol Use Disorders Identification Test; BMI, body mass index; ANOVA, analysis of variance; CEA, carcinoembryonic antigen; CRP, C-reactive protein; CT, computed tomography; DOR, diagnostic odds ratio; DSM-III, Diagnostic and Statistical Manual of Mental Disorders; ECG, electrocardiogram; EP, evoked potential; ESR, erythrocyte sedimentation rate; FNAB, fine-needle aspiration biopsy; LR, likelihood ratio; LVH, left ventricular hypertrophy; MDP, technetium-99m methylene diphosphate; MI, myocardial infarction; MS, multiple sclerosis; NBT, nitroblue tetrazolium test; NPV, negative predictive value; ns, not significant; PE, pulmonary embolism; PPi, technetium-99m(Sn) pyrophosphate; PPV, positive predictive value; RDC, research diagnostic criteria; RDOR, relative diagnostic odds ratio; SAAST, Self-Administered Alcoholism Screening Test; SBP, systolic blood pressure; SPECT, singe-photon emission computed tomography; ^{99m}Tc, technetium-99m; TRH-ST, thyrotropin-releasing hormone stimulation test; UTI, urinary tract infection.

Appendix 3

Data extraction tables: objective 2

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Adams, 1996[150] **Study design**: case series and case–control **Synthesis**: statistical	**Test**: position emission tomography **Reference standard**: CT and MRI **Condition**: head and neck, colorectal, breast, lung/solitary pulmonary nodules cancer and Alzheimer's disease	**What scale was used?** Haynes (1995)[224] and authors' own, adapted from Kent (1992)[131] and Kent (1994)[152] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence **Authors' general criteria**: *Grade A*: studies with broad generalisability to a variety of patients and no significant flaws in research methods: sample size > 70, patients drawn from clinically relevant sample with clinical symptoms completely described, diagnoses defined by an appropriate reference standard, PET studies technically of high quality and evaluated independently of references diagnosis *Grade B*: studies with narrower spectrum of generalisability, with only a few flaws that are well described: > 70 patients, more limited spectrum of patients, free of other methodological flaws that promote interaction between test results and disease determination, prospective study *Grade C*: studies with several methodological flaws: small sample size, incomplete reporting, retrospective studies of diagnostic accuracy *Grade D*: no credible reference standard, test results and determination of final diagnosis not independent, source of patient cohort could not be determined or influenced by test result, opinions without substantiating data **Quality criteria covered**: spectrum composition, reference standard, review bias, sample size	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Results for VA were presented in a table and discussed narratively
Anand, 1998[160] **Study design**: not clear **Synthesis**: narrative	**Test**: clinical assessment **Reference standard**: venography, compression ultrasonography, impedance plethysomography **Condition**: DVT	**What scale was used?** Holleman (1995)[127] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✗	The study grade was reported for 3 of the 5 studies in a table; no further reference was made to the study grading

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Attia, 1999[161] **Study design**: studies using reference standard **Synthesis**: statistical	**Test**: sensitive thyrotropin test; clinical signs and symptoms **Reference standard**: biochemical markers ± clinical features, and follow-up **Condition**: thyroid disease in acutely ill hospitalised patients	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Follow-up of normal results included vs follow-up of only abnormal results 2. > 60% of abnormal results followed up after resolution of non-thyroid illness vs < 60% 3. Criteria for diagnosis of thyroid disease clearly and explicitly stated vs implicit or clinical diagnosis **Quality criteria covered**: reference standard, work-up bias	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	VA results were reported in tabular format but not discussed in any detail. Some discussion of quality in Comment section
Bachmann, 1998[162] **Study design**: not clear **Synthesis**: statistical	**Test**: screening[162] and direct ophthalmoscopy, non-stereoscopic retinal photography[242] **Reference standard**: retinal examination by an ophthalmologist or stereoscopic retinal photography **Condition**: diabetic retinopathy	**What scale was used?** Sackett (1991)[6] **Was the QA tool designed specifically for diagnostic evaluations?** No **What type of tool was used?** Checklist	Inclusion in review: ✓ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✗	The authors state that all studies were assessed for methodological quality, but this is not referred to further. The following methodological features were used as inclusion criteria: studies in which the population studied was defined, the test was adequately described, and compared with an appropriate reference standard

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Badgett, 1996[163] **Study design**: studies with acceptable criterion standard **Synthesis**: statistical	**Test**: chest radiograph **Reference standard**: measurement of ejection fraction by non-invasive testing or by invasive pressure measurement of left ventricular preload: left ventricular end-diastolic pressure, left atrial pressure, pulmonary wedge pressure **Condition**: Left ventricular dysfunction	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Did the population include a continuous spectrum of patients that included normal patients? 2. What cut-off value of the criterion standard was used to define abnormal? 3. Were the patients clinically stable between radiographic and criterion standard assessments? 5. Was the radiographic reading blinded to the criterion standard results? **Authors' specific criteria**: 4. Were the radiographs posteroanterior films? 6. Was the radiograph interpreted by an experienced radiologist or a cardiologist? **Quality criteria covered**: time, normal defined, review bias	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✓	The implications of some quality issues were discussed
Badgett, 1997[126] **Study design**: not clear **Synthesis**: statistical	**Test**: clinical examination **Reference standard**: Left ventricular end-diastolic pressure, left atrial pressure, pulmonary capillary wedge pressure, pulmonary artery diastolic pressure **Condition**: left-sided heart failure in adults	**What scale was used?** Modified Holleman (1995)[127] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence **Authors' general criteria**: *Level 1*: independent comparison of clinical examination with suitable criterion standard among consecutive or random patients. At least 96 patients with and without a normal criterion standard *Level 2*: independent comparison of clinical examination with suitable criterion standard among consecutive or random patients *Level 3*: independent comparison of findings to a criterion standard among patients who were not consecutively or randomly chosen *Level 4*: did not have independent comparison of findings to a criterion standard	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✗	Levels of evidence were reported in summary results table, but were not considered further in the text

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Barlow, 1998[164] **Study design**: not clear **Synthesis**: narrative	**Test**: school entry medical examination **Reference standard**: follow-up **Condition**: children with health problems	**What scale was used?** Wilson and Junger criteria for screening programmes[132] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Presented in tabular format and discussed
Bastian, 1997[165] **Study design**: studies that compared clinical assessment to a reference standard **Synthesis**: statistical	**Test**: clinical signs of pregnancy **Reference standard**: urine or serum test or delivery **Condition**: pregnancy	**What scale was used?** Holleman (1995)[127] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✗	Grades of evidence were reported in tables
Bastian, 1998[166] **Study design**: not stated. Studies had to use 'appropriate controls' **Synthesis**: statistical	**Test**: home pregnancy test kits **Reference standard**: follow-up **Condition**: pregnancy	**What scale was used?** Holleman (1995)[127] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Only quality score (A/B/C) was reported

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Becker, 1996[167] **Study design**: studies that compared D-dimer results with those of objective diagnostic tests for DVT or PE **Synthesis**: narrative	**Test**: D-dimer blood measurement **Reference standard**: venogram, perfusion scan, ventilation–perfusion scan, pulmonary angiography, impedance plethysmography **Condition**: DVT or PE	**What scale was used?** Authors' own; see also Becker (1989)[206] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence and checklist **Authors' general criteria**: 1. Compared with accepted criterion standard 2. Independent interpretation of test results 3. Patient selection should be described in sufficient detail to allow a similar group of patients to be enrolled in future studies 4. Patient characteristics should be adequately described 5. Disease severity: results of test should be stratified by the extent and severity of disease 6. Patient spectrum should represent complete spectrum of patients 7. Diagnostic process: decision to perform reference standard should be made independently of the test result 8. Test descriptions sufficiently detailed to permit replication 9. Sensitivity and specificity (or the raw data to calculate these) should be given for at least one cut-off point 10. The reproducibility and interpretation of the test result should be evaluated in a setting where the test is likely to be used **Quality criteria covered**: spectrum composition, inclusion criteria, disease prevalence/severity, reference standard, test execution, reference execution, work-up bias, review bias, observer/instrument variability, appropriate results	Inclusion in review: ✗ Inclusion in primary analysis: ✓ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✓ Table: ✓ Narrative: ✓	Quality levels were used as a criteria to report studies in a table; the quality standards satisfied were also reported in this table. The methodological quality of the studies was discussed narratively and methodological factors were considered in the recommendations for future research
Bell, 1998[168] **Study design**: prospective studies **Synthesis**: narrative	**Test**: ultrasound, cancer antigen-125 **Reference standard**: histological examination of ovarian tissue for positive results and follow-up for negative results **Condition**: ovarian cancer	**What scale was used?** Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests (1996)[46] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✓ Table: ✗ Narrative: ✓	Some methodological features were discussed, not directly related to the quality assessment

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Bonis, 1997[169] **Study design**: not reported. Appears to be all cohort studies **Synthesis**: statistical	**Test**: biochemical tests **Reference standard**: histological outcome **Condition**: hepatitis C (to aid prognostic benefit)	**What scale was used?** Mulrow (1989)[45] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✓	Sensitivity analysis excluded studies with verification bias. Study quality was discussed narratively
Bradley, 1998[170] **Study design**: studies that compared a screening questionnaire with an appropriate reference standard **Synthesis**: narrative	**Test**: alcohol screening questionnaires with ≤ 10 items **Reference standard**: in-depth interviews based on standard criteria **Condition**: heavy drinking and/or alcohol abuse in general clinical populations of women in the USA	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: A. Sampling strategy for selecting participants for screening/interview was not described or only patients with positive screening results were interviewed B. Reported sensitivity was based on fewer than 20 women meeting diagnostic criteria for alcohol abuse or dependence resulting in unstable estimates, or sensitivity was not reported G. The racial and/or ethnic make-up of the study population was not explicitly stated or data for black and white patients were combined **Authors' specific criteria**: C. Items were used in more than one questionnaire, sometimes with changes in time-frame or wording, potentially limiting generalisability D. Multiple alcohol screening questionnaires were administered at one time, potentially leading to consistency response bias or other context response biases E. Screening questionnaires and comparison standards were administered by the same interviewer or at one sitting, potentially biasing questionnaire and interview responses towards higher agreement F. Criterion standards were not interview administered, potentially affecting their validities **Quality criteria covered**: spectrum composition, population recruitment, reference standard, work-up bias, sample size	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✗	Methodological limitations were noted in the main results table

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Buntinx, 1997[171] **Study design**: all papers reporting on consecutive patients with urological cancers or studies of groups of consecutive ambulatory patients with gross haematuria as the reason for encounter **Synthesis**: statistical	**Test**: macroscopic haematuria **Reference standard**: diagnosis of cancer (no further details) **Condition**: urological cancer	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: Type of data collection (prospective registration, chart review), setting, number of patients in the study, age distribution and gender ratio **Quality criteria covered**: population recruitment, spectrum composition, sample size	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✗	State that quality was assessed to inform the reader; however, results of the quality assessment do not appear to be presented
Chien, 1997[137] **Study design**: studies of pregnant women who were symptomatic or asymptomatic for preterm delivery **Synthesis**: statistical	**Test**: cervicovaginal foetal fibronectin **Reference standard**: preterm delivery before 37 or 34 weeks of gestation and delivery within 1 week after testing **Condition**: preterm delivery	**What scale was used?** Authors' own adapted from Dunn (1995),[140] Guyatt (1992)[243] and Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests (1996)[46] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Population: ideal: recruitment of consecutive women; inadequate: arbitrary recruitment; unclearly reported: information not provided 2. Diagnostic test: ideal: particular laboratory or bedside analytical test used was reported along with cut-off level for abnormal results; unclearly reported: if any of this information was missing 3a. Blinding of results: ideal: results of test were unavailable to attending physician; unclear: otherwise 3b. Completeness of follow-up: ideal: if > 90% were included in the analysis; second best: 81–90% available for analysis; inadequate: < 80% in the analysis **Authors' specific criteria**: 4. Assessment of gestational age: ideal: based on date of last menstrual period confirmed with ultrasound scan before 20 weeks of gestation; second best: in absence of menstrual date early pregnancy scans were performed to confirm gestational age; unclear: did not provide any information **Quality criteria covered**: population recruitment, normal defined, review bias, dropouts	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Study quality was discussed in the results and tabulated. Heterogeneity was investigated by stratifying on various features of methodological quality

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Conde-Agudelo, 1998[172] **Study design**: cohort **Synthesis**: statistical	**Test**: triple marker test **Reference standard**: follow-up on pregnancy outcome **Condition**: Down's syndrome	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: selection of study participants; description of technique used; estimates of sensitivity; screen-positive rate and false-positive rate; cut-off used; blinding of outcome assessments; follow-up of screened population for disease verification; accuracy estimated independently of the test threshold **Quality criteria covered**: population recruitment, test execution, work-up bias, normal defined, review bias, appropriate results, post hoc choice of threshold	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✓	Some of the implications of the quality assessment were presented in the discussion
Da Silva, 1995[173] **Study design**: studies in preterm or term infants admitted to a neonatal intensive care unit and evaluated for sepsis, in which an appropriate reference standard was used. Studies had to present information in such a way to make a 2 × 2 table **Synthesis**: narrative	**Test**: leucocyte indices and CRP **Reference standard**: neonatal infection assessed by various methods **Condition**: neonatal septicaemia	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist with quality score **Authors' general criteria**: 1. Population: ideal: consecutive infants enrolled prospectively who presented with clinical signs suggestive of sepsis admitted to neonatal intensive care unit; second best: consecutive infants who had in the past been evaluated for sepsis in neonatal intensive care unit enrolled from hospital records; worst: non-consecutive 2. Laboratory assessment: ideal: laboratory test performed on ≥ 99% of all infants; second best: performed on ≥ 90% of all infants; worst: anything else 3. Reference standard: ideal: bacterial cultures of blood, cerebrospinal fluid or urine by suprapubic or catheterisation on 99% of infants; second best: bacterial cultures on >90% of infants; worst: anything else **Quality criteria covered**: spectrum composition, population recruitment, reference standard, work-up bias	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✓	Studies were rated on a 9-point scale. Quality score was discussed briefly in the results. Methodological quality was suggested as a possible explanation for the heterogeneity between trials, but was not further investigated

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
De Bernardinis, 1999[174] **Study design**: studies with consecutive sampling and explicit definitions of severity of Ranson's criteria **Synthesis**: statistical	**Test**: Ranson's prognostic signs **Reference standard**: radiological, clinical, surgical, computed axial tomography, echotomography, post mortem and endoscopic retrograde cholangiopan-creatography **Condition**: acute pancreatitis (prediction of severity and/or prognosis)	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist with quality score **Authors' general criteria**: 1. Type of experimental design 2. Completeness of data reporting 3. Criteria for a posteriori definition of acute pancreatitis (i.e. reference standard) **Quality criteria covered**: population recruitment, reference standard, dropouts	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✓	Studies scored a maximum of 3. Association of the quality rating of the study (< 1.5/> 1.5) with the effect size was investigated
de Vries, 1996[175] **Study design**: studies had to report enough data to construct a 2×2 table and compare the test to angiography **Synthesis**: statistical	**Test**: duplex and colour-guided duplex ultrasonography **Reference standard**: contrast angiography **Population**: peripheral arterial disease	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Angiographic technique used (biplanar versus single projection) 2. Blind reading of reference standard 3. Missing observations defined as the difference between the number of segments available theoretically and the number visualised in both tests: possibility of verification bias **Quality criteria covered**: reference standard, work-up bias, review bias	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✓ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Results were presented in a table and discussed narratively. Each quality variable was entered independently into the regression analysis, none of the variables was significant

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Devous, 1998[157] **Study design**: studies with ≥ 6 patients who had a localisation-related epileptic syndrome and had at least an interictal electroencephalo-gram-documented epileptiform abnormality **Synthesis**: statistical	**Test**: SPECT brain imaging **Reference standard**: electroencephalogram and surgical outcome **Condition**: epilepsy	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Papers only assessed for one quality factor **Authors' general criterion**: Papers were assessed for blinding **Quality criterion covered**: review bias	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✓ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Results were presented in a table and discussed narratively. Blinding was included as a variable in the meta-regression
Fahey, 1995[134] **Study design**: studies that compared Pap test to histology **Synthesis**: statistical	**Test**: Pap Test: **Reference standard**: histology **Condition**: cervical precancer	**What scale was used?** Authors' own, adapted from various scales **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Independence of assessments: blind comparisons of tests 2. Selection of study participants: random, consecutive series, other or unknown 3. Point estimates of sensitivity and specificity reported and confidence intervals reported 4. Selection for disease verification: whether all participants were verified, chosen randomly or other 5. Test threshold defined: threshold separating positive and negative test results reported 6. Sampling fractions reported: if studies reported proportions of cytological positives and negatives with histological alverification 7. Accuracy estimated independently of test threshold: if accuracy was measured on an index independent of the threshold separating positive and negative test results such as ROC curve **Authors' specific criteria**: 8. Clinical use: follow-up test if prompted by findings in previous Pap test, otherwise characterised as screening 9. Technique described: if technique used to collect cervical cells was reported **Quality criteria covered**: population recruitment, work-up bias, normal defined, review bias, appropriate results, precision of results, post hoc choice of threshold	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✓ Weight the meta-analysis: ✗ Recommendations: ✓ Table: ✓ Narrative: ✓	The proportion of studies fulfilling each quality criteria was presented in a table. Summary sensitivity and specificity estimates were calculated separately for a selection of the quality criteria. Meta-regression was used to investigate the effects of these variables. Recommendations for future meta-analyses were made

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Fiellin, 2000[142] **Study design**: studies comparing a screening method to a reference standard **Synthesis**: narrative	**Test**: variety of screening methods including AUDIT and CAGE questionnaires **Reference standard**: identified diagnostic instrument **Condition**: alcohol problems in primary care	**What scale was used?** Authors' own, adapted from Reid (1995),[26] Jaeschke (1994)[35,36] and Feinstein (1985)[141] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Description of patient spectrum 2. Avoidance of work-up bias 3. Avoidance of review bias (blinding of test results) 4. Analysis of pertinent clinical subgroups **Quality criteria covered**: spectrum composition, work-up bias, review bias, subgroups	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✓ Table: ✓ Narrative: ✓	Methodological quality of studies was discussed narratively and summarised in a table. Reported on results of sensitivity analyses in the discussion
Hallan, 1997[158] **Study design**: not reported **Synthesis**: statistical	**Test**: CRP, total leucocyte count **Reference standard**: histology **Condition**: acute appendicitis	**What scale was used?** Not clear **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** No details presented **Authors' general criteria**: Authors state that "All articles were characterised and assessed for quality", but no details were provided. Some methodological characteristics were presented in discussion including: adequacy of the reference standard, avoidance of verification bias, blinded test assessment, performance of both tests in same participants, use of ROC curve analysis according to design, number of patients, prevalence and degree of disease, clinical setting, reference standard and method measuring CRP **Quality criteria covered**: spectrum composition, disease prevalence/severity, reference standard, work-up bias, review bias, sample size, post hoc choice of threshold	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✗	Not really used

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Heffner, 1995[129] **Study design**: studies whose primary purpose was to assess the value of these tests in determining the need for draining parapneumonic effusions **Synthesis**: statistical	**Test**: pleural fluid pH, lactate dehydrogenase, and glucose **Reference standard**: combination of diagnostic test results and determinations of patient outcome **Condition**: complicated parapneumonia effusions that require drainage	**What scale was used?** Authors' own, adapted from Irwig (1994)[130] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: Definition of reference standards, independence of observations, presence of verification bias **Quality criteria covered**: reference standard, work-up bias, review bias	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✓	Study quality was discussed in the results section
Heffner, 1997[147] **Study design**: studies that reported on diagnostic value of pleural effusions **Synthesis**: statistical	**Test**: pleural fluid tests **Reference standard**: clinical assessment, using explicit, objective and reproducible criteria beyond clinical judgement alone/including positive biopsy specimens **Condition**: exudative and transudative pleural effusions	**What scale was used?** Authors' own, modified from Irwig (1994)[130] and Owens (1996)[145,154] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Adequate description of sufficient reference standards 2. Independence of observations (study blinding) 3. Uniform application of reference standards 4. Assessment of generalisability 5. Cohort assembly (adequate spectrum of patients) Each component was scored as present/absent/incomplete 6. Description of biochemical testing techniques **Quality criteria covered**: spectrum composition, reference standard, test execution, work-up bias, review bias, utility of test	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✓	Study quality was discussed in the text

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Hobbs, 1997[13] **Study design**: not clear **Synthesis**: narrative	**Test**: near-patient testing (clinical chemistry, microbiology and haematology) **Reference standard**: not clear; varied for different tests **Condition**: primary care	**What scale was used?** Authors' own based on Jaeschke (1994)[35,36] and Reid (1995)[26] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist with quality score **Authors' general criteria**: 1. Was there independent blind comparison to a reference standard? 2. Did the sample include an appropriate spectrum of patients? 3. Was the reference standard performed in all patients? 4. Were the test methods described sufficiently to permit replication? 5. Are likelihood ratios quoted? 6. Would the results be reproducible in a primary care setting? 7. Would the test alter management? 8. Would the test improve patient care? 9. Are there any particular requirements or special circumstances for the use of this test?	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✓ Table: ✓ Narrative: ✓	Studies were assigned a quality score based on 5 general criteria, 1 point for each criteria, presented in data extraction tables. Results of high-quality papers were emphasised in the results section. All studies scoring 4 or more were reassessed using the Reid checklist; the validity of these studies was discussed narratively and used as recommendations for future research
Hrung, 1999[176] **Study design**: not stated **Synthesis**: statistical	**Test**: MRI **Reference standard**: excisional biopsy or mastectomy with histopathological confirmation **Condition**: primary breast cancer in patients with suspicious breast lesions	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist with quality score **Authors' general criteria**: each study was assigned a subjective quality score on a 10-point scale. Explicit details of the scale used were not given, but quality factors included: sample size, prevalence of cancer, consistent application of reference standard, verification bias evidence, prospective interpretation or blinding to reference standard, sensitivity and specificity calculable **Quality criteria covered**: disease prevalence/severity, work-up bias, review bias, sample size, appropriate results	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✓ Table: ✓ Narrative: ✓	Quality is discussed in the text and a table. Recommendations were made on how future studies should be conducted

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Huicho, 1996[128] **Study design**: studies in which patients were investigated with faecal screening and culture which provided enough data to calculate sensitivity and specificity **Synthesis**: statistical	**Test**: faecal leucocytes, faecal occult blood, faecal lactoferrin and combination of faecal leucocytes with clinical data **Reference standard**: stool culture **Condition**: inflammatory bacterial diarrhoea	**What scale was used?** Mulrow (1989),[45] (modified) **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist with quality score	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	The results of the quality assessment were presented in a table and study quality was discussed in the discussion section
Kearon, 1998[177] **Study design**: prospective cohort studies **Synthesis**: statistical	**Test**: non-invasive approaches: impedance plethysmography and venous ultrasonography alone or in combination with other tests **Reference standard**: venography **Condition**: DVT	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: to be included in the review: 1. All patients had to receive the reference standard test 2. The 2 test must have been evaluated independently (blinded) 3. Consecutive patients must have been studied 4. The study must have been prospective 5. At least 50 patients had to be studied **Quality criteria covered**: population recruitment, work-up bias, review bias, sample size	Inclusion in review: ✓ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✗	Recommendations made were graded by quality (A–C)
Koelemay, 1996[178] **Study design**: reference standard studies reporting sensitivity and specificity **Synthesis**: statistical	**Test**: duplex ultrasonography **Reference standard**: angiography **Condition**: peripheral arterial occlusive disease	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence **Authors' general criteria**: 1. Clear definition of the study population 2. Clear description of duplex scanning technique 3. Series of consecutive patients 4. Prospective study 5. Predefined test criteria and independent assessment of both tests **Quality criteria covered**: spectrum composition, population recruitment, reference standard, test execution, normal defined, review bias	Inclusion in review: ✗ Inclusion in primary analysis: ✓ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Studies satisfying all criteria were graded level 1, studies satisfying at least the 2 essential criteria (1 and 2) were graded level 2, all others were graded level 3. Only level 1 and selected level 2 studies were included in the quantitative analysis. Study level was presented in a table and discussed in the text

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Koumans, 1998[148] **Study design**: not stated **Synthesis**: statistical	**Test**: non-culture tests (nucleic acid hybridisation or amplification tests) **Reference standard**: culture **Condition**: *Neisseria* gonorrhoea	**What scale was used?** Authors' own, adapted from Owens (1996)[145,154] and Irwig (1994)[130] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist with quality score **Authors' general criteria**: 1. Sample size 2. Test quality (sufficient detail of test and reference standard) 3. Blinding of investigators to test results and clinical information 4. Clinical description of sample (whether description of study sample was complete) 5. Assembly of population (adequate spectrum; sufficient description of assembly) 6. Consistent application of reference standard to diseased and non-diseased population **Authors' specific criteria**: if gender of participants was not described or there were < 5 culture-positive participants, a sample size score was not assigned and performance results were neither abstracted nor combined, unless specimens were taken from pharynx or rectum **Quality criteria covered**: spectrum composition, population recruitment, test execution, reference execution, work-up bias, review bias, sample size	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	VA results were itemised in tabular format and discussed
Lacasse, 1999[136] **Study design**: series of consecutive patients where ≥ 90% underwent the reference standard **Synthesis**: statistical	**Test**: transthoracic needle aspiration biopsy **Reference standard**: resection specimen, biopsy procedures of an adjacent site with tumour involvement, long-term follow-up or culture **Condition**: localised pulmonary lesions	**What scale was used?** Authors' own, based on Jaeschke (1994)[35,36] and Begg (1988)[135] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Work-up bias 2. Review bias 3. Test review bias 4. Prospective or retrospective evaluation (NB. Criteria 1 and 2 were inclusion criteria for the review)	Inclusion in review: ✓ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	VA results were discussed and some detail was presented in a table. Included only consecutive series of patients (to avoid work-up bias and ensure adequate patient spectrum). Required ≥ 90% of patients to undergo reference standard (to avoid review bias)

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Lederle, 1999[179] **Study design**: series with ≥ 10 patients **Synthesis**: statistical	**Test**: physical examination (abdominal palpation) **Reference standard**: ultrasound **Condition**: abdominal aortic aneurysm	**What scale was used?** Holleman (1995)[127] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✓	VA results were discussed narratively
Liedberg, 1996[180] **Study design**: studies comparing the test with a reference standard **Synthesis**: narrative	**Test**: arthrography, CT and MRI **Reference standard**: surgery, clinical + imaging, cryosection, macroscopy, arthrography **Condition**: temporomandibular joint disorder	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: not reported, only state that studies were weighted according to quality **Quality criteria covered**: not reported	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✓	Studies were weighted according to their relative quality. Only those studies with a disease prevalence between 32 and 78% were considered. When the reference standard consisted of another imaging method, the study was discarded
Littenberg, 1995[181] **Study design**: studies that estimated the diagnostic accuracy of SPECT in humans with low back pain **Synthesis**: narrative	**Test**: SPECT bone imaging **Reference standard**: medical records/telephone conversation, biopsy, CT, MRI, follow-up planar bone scan, follow-up plain films or clinical examination, surgery. **Condition**: low back pain	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: not a formal quality assessment, but data were extracted on the following quality-related features: 1. Number of participants 2. Patient sources (referral bias) 3. Presence of a reference standard (including surgical findings or long-term clinical follow-up) 4. Blinding **Quality criteria covered**: population recruitment, reference standard, review bias, sample size	Inclusion in review: ✗ Inclusion in primary analysis: ✓ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✓ Table: ✗ Narrative: ✓	Only studies that compared SPECT to a reference standard and provided sufficient information to construct a 2 × 2 table were reported in the primary analysis. The authors discuss study quality and report on the characteristics of an ideal study

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Loy, 1996[125] **Study design**: studies that compared the accuracy of different tests **Synthesis**: statistical	**Test**: commercial serological kits (ELISA and latex agglutination) **Reference standard**: culture, histology, urease testing on biopsy and other **Condition**: *helicobacter pylori*	**What scale was used?** Authors' own, adapted from Jaeschke (1994)[35,36] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: unbiased sample selection, appropriate blinding, reference standard used	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✓ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Details were presented in a table and discussed in the text. Quality variables were included individually in a regression analysis
Mayer, 1997[182] **Study design**: studies that compared dermatoscopy with another method of clinical diagnosis that used appropriate reference standard **Synthesis**: narrative	**Test**: dermatoscopy and clinical diagnosis **Reference standard**: excision biopsy with histopathological examination **Condition**: malignant melanoma	**What scale was used?** Sackett (1991)[6] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Some methodological characteristics were presented in a table and discussed in the text. The authors state that quality assessment was hampered by a lack of information. Only one study provided sufficient detail of the methods
McGee, 1999[183] **Study design**: studies that looked at patients presenting at emergency departments with suspected hypovolaemia. Studies of healthy volunteers also included **Synthesis**: statistical	**Test**: physical/bedside diagnosis **Reference standard**: various **Condition**: hypovolaemia in adults	**What scale was used?** Not clear whether specific scale; appears to be authors' own criteria listed as footnote to table **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence **Authors' general criteria**: graded as A, B or C: A, an independent blind comparison of a defined physical sign with an acceptable reference standard in more than 50 consecutive patients; B, same as A but fewer than 50 consecutive patients; C, all other studies **Quality criteria covered**: reference standard, review bias, sample size	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✗	The study grades were reported in a table but were not discussed further or used in the analysis

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Metlay, 1997[184] **Study design**: original studies of the accuracy or precision of the history and/or physical examination in diagnosing pneumonia **Synthesis**: narrative	**Test**: history and physical examination **Reference standard**: new infiltrate on chest radiograph **Condition**: community-acquired pneumonia	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence **Authors' general criteria**: Level 1: primary, prospective study of the accuracy/precision of clinical examination. Independent blind comparisons of clinical findings with reference standards among a large number (> 50) of consecutive patients Level 2: same as level 1 with smaller number of patients (10–50) Level 3: included non-consecutive patients, generally selected because of their definitive results for the findings under study, or a non-blinded comparison with reference standard Level 4: included studies with uncertain reference standards or poorly defined study population **Quality criteria covered**: population recruitment, reference standard, work-up bias, review bias, sample size	Inclusion in review: ✗ Inclusion in primary analysis: ✓ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✗	Only studies of level 1 quality were considered for the main analyses and tables
Mol, 1997[185] **Study design**: studies comparing the two tests; studies had to report sufficient data to construct a 2 × 2 table **Synthesis**: statistical	**Test**: *Chlamydia* antibody titres **Reference standard**: laparoscopy **Condition**: tubal pathology in subfertile patients	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Method of inclusion (consecutive vs non-consecutive) 2. Type of verification used to define criteria for tubal disease 3. Independent performance of reference standard **Quality criteria covered**: population recruitment, reference standard, review bias	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✓ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✗	Quality variables were included in a logistic regression analysis

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Mol, 1998[186] **Study design**: cohort and case–control **Synthesis**: statistical	**Test**: single serum progesterone measurement **Reference standard**: various: surgery, histology, sonography, delivery, dilation and curettage, falling human chorionic gonadotropin Condition: Ectopic pregnancy; Pregnancy failure	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Sampling (consecutive or other) 2. Data collection (prospective vs retrospective) 3. Study design (cohort vs case–control) **Quality criterion covered**: population recruitment	Inclusion in review: ✗ Inclusion in primary analysis: ✓ Sensitivity analyses: ✗ In regression analysis: ✓ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Results of the quality assessment were presented in a table and discussed in the text. The quality assessment variables were included in a regression analysis. Further analyses were limited to cohort studies
Mol, 1998[187] **Study design**: cohort and case–control **Synthesis**: statistical	**Test**: cancer antigen-125 serum **Reference standard**: laporoscopy **Condition**: endometriosis	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Study design (cohort vs case–control) 2. Sampling (consecutive or other) 3. Data collection (prospective vs retrospective) 4. Blinding 5. Verification bias 6. Establishment of final diagnosis **Quality criteria covered**: population recruitment, reference standard, work-up bias, review bias	Inclusion in review: ✗ Inclusion in primary analysis: ✓ Sensitivity analyses: ✗ In regression analysis: ✓ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Results of the quality assessment were presented in a table and discussed in the text. The quality assessment variables were included in a regression analysis. Further analyses were limited to cohort studies

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Mullins, 2000[156] **Study design**: studies that compared spiral volumetric CT to clinical reference standard **Synthesis**: narrative	**Test**: spiral volumetric CT **Reference standard**: pulmonary arteriogram or another clinical reference standard, e.g. V/Q scan **Condition**: pulmonary embolism	**What scale was used?** Authors' own, adapted from Ransohoff (1978)[66] and Philbrick (1980)[155] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Was the test technique clearly described? 2. Did the authors provide clear criteria for positive/negative test results? 3. Were results interpreted blindly? 4. Was the reliability of the test assessed by having some patients undergo repeated comparisons of both tests? 5. Was the selection of patients adequately described? 6. Were the patients adequately described? 7. Were eligible patients who were not enrolled described sufficiently? 8. Was the extent of disease described in sufficient detail to allow stratification of results by location or severity of disease? 9. Were non-diseased results reported? 10. Were patients referred to the reference standard regardless of the results of either test? 11. Were results of the 2 tests interpreted independently? **Quality criteria covered**: spectrum composition, population recruitment, disease prevalence/severity, test execution, work-up bias, normal defined, review bias, observer/instrument variability	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	The authors report which criteria were fulfilled in a table and quality of studies was discussed

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Nuovo, 1997[149] **Study design**: studies in which reference standard was performed in all patients **Synthesis**: narrative	**Test**: cerviography **Reference standard**: colposcopy **Condition**: cervical cancer	**What scale was used?** Authors' own, based on Reid (1995),[26] Jaeschke (1994)[35,36] and Irwig (1994)[130] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors general criteria**: 1. Did the patient sample include an appropriate spectrum of patients? 2. Were the study setting and the filter through which patients passed adequately described? 3. Were reproducibility and interpretation of test results determined? 4. Was the term normal defined sensibly? 5. If the test is advocated as part of a cluster or sequence of tests, was its contribution to the overall validity of the cluster or sequence determined? 6. Were tactics for carrying out the test described in sufficient detail to permit their exact replication? 7. Was the utility of the test determined? 8. Are the results applicable in primary care patients? 9. Will the results lead to a change in management? **Quality criteria covered**: spectrum composition, population recruitment, test execution, work-up bias, normal defined, observer/instrument variation, utility of test	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✓ Table: ✓ Narrative: ✓	The authors discuss the methodological limitations of the studies in light of the quality assessment scales and state that future high-quality research is required. Results were summarised in a table

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Owens, 1996[145,154] **Study design**: studies in which DNA amplification by PCR was done on peripheral blood mononuclear cells from adults **Synthesis**: statistical	**Test**: PCR **Reference standard**: enzyme immunoassay followed by confirmatory Western blot analysis; viral culture: serial testing or follow-up **Condition**: HIV in adults and infants	**What scale was used?** Authors' own, based on Hoffman (1991),[133] Kent (1992)[131,151] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. PCR test quality: was the test described in sufficient detail to be reproducible? 2. Reference standard quality 3. Application of reference standard: was the test applied consistently in the diseased and non-diseased individuals? 4. Blinding: were investigators blinded to all other test and clinical information? 5. Clinical description: was the study population described adequately? 6. Cohort assembly: was the spectrum of patients adequate? 7. Sample size: was the sample size adequate (> 30 in diseased and non-disease group) **Quality criteria covered**: spectrum composition, population recruitment, reference standard, test execution, work-up bias, review bias, sample size	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✓ Table: ✓ Narrative: ✓	Results of validity assessment were tabulated, graphed and discussed. The authors make detailed suggestions for improving future study designs
Pearl, 1996[188] **Study design**: prospective studies **Synthesis**: narrative	**Test**: ultrasonography **Reference standard**: diagnostic peritoneal lavage, CT or laporotomy **Condition**: blunt	**What scale was used?** Kent (1992)[131] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence and checklist	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✓ Table: ✓ Narrative: ✓	Results of quality assessment were presented in a table and the text. Quality was also discussed narratively. The authors make recommendations for further research based on study quality

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Pollitt, 1997[189] **Study design**: not stated **Synthesis**: narrative	**Test**: tandem mass spectrometry **Reference standard**: various **Condition**: inborn errors of metabolism	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence **Authors' general criteria**: Level 1: data obtained from screening programmes in UK population or similar Level 2: data from systematic studies other than from whole population screening Level 3: estimated from the known biochemistry of the condition: **Quality criterion covered**: population recruitment	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✗	Studies were graded according to level of evidence; not really discussed in the results
Rao, 1995[146] **Study design**: studies that provided sufficient information for construction of a 2 × 2 table **Synthesis**: statistical	**Test**: antineutrophil cytoplasmic antibody **Reference standard**: ear, nose, throat, lung and kidney staging system biopsy, fauci criteria (clinicopathological) and the American College of Rheumatology criteria **Condition**: Wegener granulomatosis	**What scale was used?** Authors' own, based on Sackett (1991)[6] and Irwig (1994)[130] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist with quality score **Authors' general criteria**: Study design: cohort, 3 points; case–control, 2; case series, 1; uncertain, 0 Data collection: prospective, 3; retrospective–prospective, 2; retrospective, 1; uncertain, 0 Selection of patients: random or consecutive, 3; other, 0 Selection of controls: disease could be ruled out in all, 3; in > 50%, 2; in < 50%, cannot tell, 1; could not be ruled out, 0 Were biopsy results read blinded to diagnosis under consideration? yes, 1; no, 0 **Quality criteria covered**: population recruitment, review bias	Inclusion in review: ✓ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✓	Only studies with high-quality reporting of methods were included in the review. Quality score was reported narratively. Results were stratified on the basis of quality score
Rappeport, 1996[159] **Study design**: studies comparing MRI and arthroscopy **Synthesis**: narrative	**Test**: MRI **Reference standard**: arthroscopy **Condition**: Knee injury: acute or chronic complaint of the knee joint (including injury or degenerative disease)	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Discuss how study compares to 'ideal study' **Authors' general criteria**: The authors state that an ideal study would be "double blind, prospective study, in which tests are performed independently, in a non-selective group of independent patients" **Quality criteria covered**: population recruitment, work-up bias, review bias	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✗	The authors discuss how many studies met the ideal standard. Results for the higher quality studies are discussed separately. Study results were presented in a table

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Reed, 1996[190] **Study design**: studies had to report sufficient information to generate a 2 × 2 table and compare sputum stain with an independent reference standard for Inclusion in review **Synthesis**: statistical	**Test**: sputum Gram stain **Reference standard**: culture **Condition**: community-acquired pneumococcal pneumonia	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Explicit inclusion and exclusion criteria 2. Assessment of inter- and intraobserver variability 3. Training of test interpreter 4. Whether assessment of Gram's stain test characteristic was a specific objective of the study 5. General clarity of study **Quality criteria covered**: inclusion criteria, observer/instrument variability, objective	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Quality assessment results were presented in a table and discussed in the text
Selker, 1997[191] **Study design**: clinical trials in the emergency department setting: studies of test sensitivity and specificity and studies of the clinical impact of the test's actual use **Synthesis**: statistical	**Test**: diagnostic technologies used in the emergency department **Reference standard**: various **Condition**: acute cardiac ischaemia	**What scale was used?** Mulrow (1989)[45] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✗	Study gradings were shown in summary tables, but quality was not discussed further
Spencer-Green, 1997[192] **Study design**: case–control studies **Synthesis**: narrative	**Test**: antibody tests (anti-centromere and anti-Scle-70) **Reference standard**: authors' classification of systemic sclerosis (previously published) **Condition**: systemic sclerosis	**What scale was used?** Mulrow (1989)[45] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist	Inclusion in review: ✓ Inclusion in primary analysis: ✓ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Study ratings were discussed narratively and presented in a histogram. Looked at study rating as a possible source of heterogeneity; it was not found to be associated

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Swart, 1995[193] **Study design**: all retrospective **Synthesis**: statistical	**Test**: hysterosalpingography **Reference standard**: laporoscopy with chromopertubation **Condition**: tubal pathology (absence of tubal patency and presence of peritubal adhesions)	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist **Authors' general criteria**: 1. Setting (academic or non-academic) 2. Independent interpretation of reference standard 3. Use of criteria for interpretation of index test 4. Time between test and reference standard 5. Number of patients 6. Disease prevalence (< or ≥ 35%) **Authors' specific criteria**: 7. Contrast use (oil vs water) 8. Presence of spasmolyticum (yes or no) **Quality criteria covered**: population recruitment, disease prevalence/severity, time, normal defined, review bias, sample size	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Some subgroup analyses relate to quality assessment, but not explicitly stated. Results were presented in a table and discussed in the text
Tugwell, 1997[194] **Study design**: not reported **Synthesis**: statistical	**Test**: laboratory diagnosis, including culture, ELISA, Western blot **Reference standard**: criteria for confirmed infection **Condition**: Lyme disease	**What scale was used?** Irwig (1994)[130] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist	Inclusion in review: ✓ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✗	Included studies had to provide a clear statement on the test of interest, a description of the study characteristics that used a design that permitted the calculation of sensitivity and specificity, reproducible information on the sampling and clinical details of patients with the disease of interest and on controls, and reproducible information on the reference standard

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
van den Hoogen, 1995[144] **Study design**: studies with >10 patients **Synthesis**: narrative	**Test**: history, physical examination, and ESR **Reference standard**: anatomical findings at surgery; overall clinical impression after diagnostic imagery or New York criteria (for ankylosing spondylitis) **Condition**: low back pain resulting from radiculapathy, vertebral cancer metastasis, ankylosing spondylitis	**What scale was used?** Authors' own, adapted from Hoffmann (1991),[133] Kent (1992)[131] and Mulrow (1989)[45] **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist with quality score **Authors' general criteria**: 1. Technical quality of the index test 2. Technical quality of the reference standard 3. Application of the reference standard 4. Independence of interpretation 5. Clinical description 6. Study population (prospective enrolment, adequate patient spectrum, explicit inclusion criteria) 7. Study population (inclusion of both diseased and non-diseased participants) 8. Sample size 9. Data presentation **Quality criteria covered**: spectrum composition, inclusion criteria, population recruitment, test execution, reference standard execution, work-up bias, review bias, sample size, data table	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✓ Table: ✓ Narrative: ✓	Study scores for each criteria were presented in a table; the methodological quality of studies was discussed narratively and used to make recommendations for future research
van Tulder, 1997[139] **Study design**: studies that included a study population with and without low back pain and for whom at least one of the diagnostic tests was plain radiography **Synthesis**: narrative	**Test**: spinal radiographs **Reference standard**: various **Condition**: low back pain	**What scale was used?** Authors' own, adapted from various checklists **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Checklist with quality score **Authors' general criteria**: Study population: selection of study population, description of inclusion and exclusion criteria, description of potential confounders Blinded assessment: blinded assessment of radiographs, blinded assessment of low back pain status Analysis and data presentation: no missing values or description of missing values, presentation or reconstruction of 2 × 2 tables, control for confounders Assessment of radiographs: description of technique and equipment, definition of normal and abnormal result, reproducibility of test interpretation Assessment of low back pain status: appropriate test for low pack pain, same test applied to all participants, adequate follow-up period **Quality criteria covered**: inclusion criteria, population recruitment, reference standard, test execution, work-up bias, normal defined, review bias, observer/instrument variability, dropouts, data table	Inclusion in review: ✓ Inclusion in primary analysis: ✓ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Studies were given a percentage score and the methodological quality of studies was discussed. Internal and external validity (quality of reporting) were considered separately. Only results of studies of acceptable or good methodological quality were presented in the main results table. Results of the quality assessment were tabulated

continued

Study details	Test details	Quality assessment scale details	How was the quality assessment used in the review?	Further details of quality assessment scale use
Wells, 1995[195] **Study design**: prospective and retrospective **Synthesis**: statistical	**Test**: ultrasound **Reference standard**: standard contrast venography **Condition**: DVT	**What scale was used?** Authors' own **Was the QA tool designed specifically for diagnostic evaluations?** Yes **What type of tool was used?** Levels of evidence and checklist **Authors general criteria**: 1. Previous establishment of objective criteria for normal and abnormal venographic and ultrasonographic results 2. Independent comparison of index test with reference standard by investigators blinded to the other test result 3. Prospective evaluation of consecutive eligible patients **Quality criteria covered**: population recruitment, reference standard, work-up bias, normal defined, review bias	Inclusion in review: ✗ Inclusion in primary analysis: ✗ Sensitivity analyses: ✓ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✓ Narrative: ✓	Studies were classified as level 1 or 2 according to quality and comparisons were made between the results of these studies. Study level was presented in a table and discussed in the text
Whited, 1998[196] **Study design**: not stated **Synthesis**: narrative	**Test**: clinical examination **Reference standard**: histopathological examination of excised tissue **Condition**: melanoma	**What scale was used?** Holleman (1995)[127] **What type of tool was used?** Checklist **Was the QA tool designed specifically for diagnostic evaluations?** Yes	Inclusion in review: ✓ Inclusion in primary analysis: ✗ Sensitivity analyses: ✗ In regression analysis: ✗ Weight the meta-analysis: ✗ Recommendations: ✗ Table: ✗ Narrative: ✓	Studies were included if they were level C or above

QA, quality assessment; VA, validity assessment; DVT, deep venous thrombosis; ELISA, enzyme-linked immunosorbent assay; PCR, polymerase chain reaction.

Appendix 4

Data extraction tables: objective 3

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Anon, 1981[47] **Aim**: guidelines for interpreting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: did the patient sample include an appropriate spectrum of mild and severe, treated and untreated disease, plus individuals with different but commonly confused disorders? **Population recruitment**: was the setting for the study as well as the filter through which study patients passed, adequately described? The article should provide enough information about the study site and patient selection filter to permit calculation of the test's likely predictive value in other situations of interest. Description of control subjects should also be adequate	**Appropriate reference standard**: was there a comparison with a 'reference standard' of diagnosis? The reference standard should give a definitive diagnosis attained by biopsy, surgery, autopsy, long-term follow-up or another acknowledged standard **Test execution**: were the tactics for carrying out the test described in sufficient detail to permit their exact replication? This description should cover patient issues as well as the mechanics of performing the test and interpreting its results **Normal defined**: was the term 'normal' defined sensibly? If the article uses the word 'normal' its authors should say what is meant by it. You should satisfy yourself that their definition is clinically sensible	**Review bias**: was the comparison 'blind'? Patients should have undergone the diagnostic test and the test should have been interpreted by clinicians who were blind to whether a given patient really had the disease **Observer/instrument variation**: was the reproducibility of the test results (precessions) and its interpretation (observer variation) determined? The description of a diagnostic test ought to tell readers how reproducible they can expect the test results to be; this is especially true when expertise is required in performing the test or in interpretation	**Appropriate results**: if the test is advocated as part of a cluster or sequence of tests, was its contribution to the overall validity of the cluster or sequence determined? In many conditions an individual diagnostic test examines only one of several manifestations of the underlying disorder. Any single component of a cluster of diagnostic tests should be evaluated in the context of its clinical use **Data table**: if you do not find or cannot construct a 2×2 table from the results paper it is probably not worth reading further **Utility of test**: was the 'utility' of the test determined? Authors should go beyond the issues of accuracy, precision, etc., to explore the long-term consequences of their use of the diagnostic test. In addition to describing what happened to patients correctly diagnosed they should also describe the fate of the patients who had false-positive results and those with false-negative results. When the execution of a test requires a delay in the initiation of therapy the consequences of this delay should be described		**How were items chosen for inclusion on the scale?** not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Arrive, 2000[208] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: *Spectrum of patients*: age and gender distribution, summary of clinical symptoms at presentation and an indication of disease severity had to be provided **Inclusion criteria**: *Inclusion criteria*: a precise plan for population inclusion had to be described *Exclusion criteria*: both explicit exclusion criteria and the number of excluded subjects had to be provided **Population recruitment**: *Study design*: consecutive/non-consecutive recruitment, prospective or retrospective	**Appropriate reference standard**: *Reference standard*: had to include clear definition of the reference standard and that the reference standard be an accurate method for assessing the presence of disease **Test execution**: *Method of analysis*: details on performing the examination and interpreting the results had to be provided, including explicit descriptions of the techniques used and of the analysis process **Work-up bias**: *Avoidance of verification or work-up bias*: when all patients underwent both the radiological examination under evaluation and the reference standard procedure, or in the case of an invasive reference standard, when a validated adjustment to correct for verification bias was used **Incorporation bias**: when the radiological examination under investigation was actually incorporated into the evidence used as the reference standard **Normal defined**: *Analysis criteria*: had to be	**Review bias**: *Avoidance of diagnostic review bias*: statement was required that the results of the reference standard were interpreted independently of the results of radiological examination being investigated *Avoidance of test review bias*: statement was required indicating that the radiological examination results were interpreted without knowledge of the reference standard results **Observer/instrument variation**: *Intraobserver reliability*: if an appropriate statistical test was used for evaluation of intraobserver reliability *Interobserver reliability*: if an appropriate statistical test was used for evaluation of interobserver reliability	**Appropriate results**: *Statistical analysis*: when all appropriate statistical analyses were precisely described and performed **Indeterminate test results**: *Indeterminate examination results*: required statements regarding the existence and frequency of indeterminate examination results and the manner in which indeterminate examination results were accounted for in the estimation of accuracy	**Objective**: adequate definition of the purpose of the study	**How were items chosen for inclusion on the scale?** The scale was based on a list of methodological standard that can be applied to any clinical study of radiological examination evaluation. The methodological standards were compiled from established resources for assessing the methodological quality of investigational design for studies in clinical research and from literature related to biases commonly observed in radiological research. Study design had to be described, all other standards scored as 'yes', 'partially' or 'no' **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** not reported **Level of inter-rater reliability**: agreement in ratings between observers for the 15 standards was high, with kappa values of 0.9–1.0 for 8 standards, 0.8–0.9 for 3 standards and 0.7–0.8 for the remaining 4 standards. Agreement

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
		described in adequate detail for both positive and negative results and when analysis criteria had been defined before the study commenced				in ratings for the composite quality index was high, with an intraclass correlation coefficient of 0.91 (95% CI 0.87 to 0.94) **Topic area**: radiology
Beam, 1991[29] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: modified tool	**Spectrum composition**: were patients 'profiled'? (a) Quantitatively only, (b) qualitatively only, (c) both, (d) neither. Were groups statistically compared with respect to above profile? (yes/no) **Inclusion criteria**: reasons for exclusions of patients: (a) given, (b) not given, (c) none excluded **Population recruitment**: how were patients selected? (a) Consecutive stream, (b) true random sample, (c) other, (d) cannot tell	**Appropriate reference standard**: presentation of a standard reference: (a) tissue diagnosis only, (b) other imaging procedure only, (c) both, (d) history, (e) none given, (f) not applicable	**Review bias (test and diagnosis)**: random order of imaging when comparing MRI with another procedure (yes/no) **Clinical review bias**: blinding of interpreter with respect to clinical history or other test result (yes/no/not applicable) **Observer/instrument variation**: measurement of interobserver/instrument variation in reading images (yes/no)	**Appropriate results**: substantiation of assumptions behind statistical analyses (yes/no/no statistical analysis done) **Precision of results**: estimate of variability of parameter estimates (yes/no/no parameters estimated)	**Sample size**: sample sizes given (yes/no) **Protocol**: evidence of research planning: (a) institutional review board only, (b) protocol only, (c) both, (d) neither	**How were items chosen for inclusion on the scale?** Modified version of Cooper (1988)[30] (the objective criteria were retained) **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** not reported **Level of inter-rater reliability**: not reported **Topic area**: MRI

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Becker, 1989[206] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: systematic review; authors' own tool	**Spectrum composition**: *Identification of groups selected for study*: to clarify the clinical spectrum of patients tested, the age and gender of the patients had to be reported, along with a brief summary of the major clinical characteristics of the patients tested *Analysis of the anatomical extent of disease*: a description of the criteria for the venographic diagnosis of DVT was required, together with a report of the most proximal extension of the DVT and the results of ultrasonography for at least one anatomical subgroup *Analysis of conditions that mimic DVT*: a summary of non-DVT diagnoses had to be reported along with the results of the ultrasound tests for each diagnosis **Population recruitment**: *Identification of groups selected for study*: the method of selecting patients had to be described in sufficient detail to allow a similar group of patients to be selected if the study were repeated	**Test execution**: *Description of ultrasonography technique*: explicit description of the ultrasonographic technique used was required to allow others to perform the examination in a similar way **Verification bias**: *Avoidance of work-up bias*: a study design had to be used that committed patients to receive both the ultrasound and venogram before the ultrasound was performed **Normal defined**: *Description of ultrasonography technique*: required a clear description of the criteria for positive and negative examinations used in the study	**Review bias**: *Avoidance of diagnostic review bias*: a statement that the interpreters of the venogram did not know that the ultrasound result was required *Avoidance of test review bias*: a statement that the interpreters of the ultrasound were blinded to the results of the venogram was required **Observer/instrument variation**: *Assessment of test reliability*: required repeated ultrasonography of some of the patients by at least one additional examiner and blinding of the examiners to each other's findings when interpreting the test result			**How were items chosen for inclusion on the scale?** The standards were adapted in large part from Ransohoff (1978)[66] and Philbrick (1980)[155]; no further details **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** not reported **Level of inter-rater reliability**: not reported **Topic area**: ultrasound for DVT

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Black, 1990[210] **Aim**: assess study quality **Type of scale**: Checklist **Source of tool**: original tool	**Spectrum composition**: are the patients representative of those who are ordinarily tested? **Population recruitment**: how are the patients selected? Ideally patients are described in such a fashion that similar patients can be identified easily for future application of the study results. Retrospective vs prospective selection	**Appropriate reference standard**: what is the reference standard for diagnosis? Is it appropriate? **Test execution**: how is the test performed and interpreted? **Verification bias**: is the reference standard applied uniformly? **Normal defined**: are the interpretation criteria well defined and reproducible?	**Review bias (diagnostic and test)**: are the radiologists blinded from the final diagnosis and is the final diagnostician blinded from the radiological interpretation?	**Appropriate results**: in a comparison study are the tests evaluated fairly? Are spectrum of disease and important covariates, such as comorbidity, age, gender and body habitus, accounted for in tabular presentation of state? Is the statistical analysis clearly described and appropriate? **Data table**: how is accuracy reported?		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: radiology

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Bruns, 2000[222] **Aim**: guidelines for reporting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: patient care setting: describe and provide summary of factors, especially other tests, that channelled patients to have the test under evaluation. Methods to avoid spectrum bias (e.g. consecutive series, statistically selected random sample, stratified sample) and to define spectrum of disease. Design features aimed at ensuring comparability with other studies. Report demographics of subjects and their clinical characteristics to include the presence or absence of disease and spectrum of disease **Inclusion criteria**: criteria for inclusion and exclusion of subjects, especially regarding the results of any other tests and the criteria used in the interpretation of those tests, consent procedures and approvals of study. Report numbers of subjects who were excluded, reasons for exclusions, and their timing **Population recruitment**: state study design used. Report inclusive date of accrual of subjects	**Test execution**: *Methods (and references) for evaluated tests*: references should include studies that validate the analytical performance of the tests; when no such studies have been published, provide key information **Description of reference standard execution**: *Methods (and references) for reference standard*: references should include studies that validate the analytical performance of the tests; when no such studies have been published, provide key information. When an outcome is used as the criterion standard, indicate duration and methods of follow-up **Verification bias**: methods to avoid verification bias (usually by application of reference standard to all subjects) or to deal with its consequences **Normal defined**: cut-offs used for quantitative tests and how they were determined; subjective criteria for qualitative tests **Treatment paradox**: (for prognostic tests) indicate whether the criterion standard or the evaluated tests influenced therapy	**Review bias (test and diagnostic)**: indicate the blinding of those performing the evaluated test and criterion standard test to avoid reviewer bias. When multiple tests are to be evaluated indicate whether the performance of each test was without knowledge of the results of the others **Observer/instrument variation**: report data on reproducibility of evaluated tests: intra- and interobserver/ instrument variation	**Appropriate results**: Methods (and references) for statistical analysis including steps to deal with potential for diagnostic accuracy to be overestimated when diagnostic rules are constructed by use of statistical modelling or by examination of more than one cut-off value for continuous variables, repeated or serial measures and outliers. Report deviations (e.g. loss to follow-up) from study protocol and reasons. Report repository where original data may be obtained (e.g. for use in systematic reviews) **Precision of results**: report measures of diagnostic accuracy of tests and confidence intervals **Indeterminate test results**: report number of indeterminate test results and their use in further data analysis **Data table**: appropriate tabulation of key results (e.g. 2 × 2 table)	**Sample size**: planned sample size: report statistical power and resource consideration; report sample size achieved	**How were items chosen for inclusion on the scale?** Examined published reports on shortcomings of studies of diagnostic accuracy, prepared an initial draft checklist to address common errors and presented it at a meeting of editors. After incorporation of comments from editors, published revised version[245] for comment from readers. Circulated copies of the draft to methodologists and others interested in evidence-based medicine. Updated the checklist with input from these sources **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** See above for details **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Cochrane, 1996[46] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: *Spectrum of disease*: e.g. cancer stage distribution in those categorised as 'diseased' by the reference standard *Spectrum of non-disease*: e.g. case-mix in those categorised as 'non-diseased' by the reference standard *Setting*: primary care, tertiary care, outpatients, inpatients, etc. Duration of illness before testing; previous tests/ referral filter, i.e. to what clinical (including previous test) information is the test being evaluated; co-morbid conditions; demographic information: age, gender or other variables that may act as proxies of the above **Population recruitment**: prospective or retrospective design **Disease prevalence/ severity**: disease prevalence/severity	**Appropriate reference standard**: was the test compared with a valid reference standard? Categorise studies by type of reference standard used **Test execution**: categories of how the test was done, e.g. types of biochemical methods **Verification bias**: was the choice of patients who were assessed by the reference standard independent of the test's result (avoidance of verification bias)? 1. Reference standard measured in consecutive people 2. Random sample of consecutive people 3. Random sample of people who are positive or negative by the test and adjusting for different sampling fractions 4. Measuring the test in random samples of people within groups defined by the reference standard as diseased or non-diseased Were tests compared in a valid design? 1. All tests done independently (i.e. blind to the results of the other tests) on each person (most valid) 2. Different tests done on randomly allocated individuals 3. All tests done on each person but not assessed independently 4. Different tests done on different individuals, not randomly allocated (least valid) **Normal defined**: state the explicit threshold used **Treatment paradox**: was the reference standard measured before any interventions were started with knowledge of test results (avoidance of treatment paradox)?	**Review bias (test and diagnostic)**: were the test and reference standard measured independently (blind) of each other? 1. Test measured independently of reference standard and reference standard independently of test (most valid) 2. Test measured independently of reference standard but not vice versa 3. Reference standard measured independently of test but not vice versa 4. Test and reference standard not measured independently of each other (least valid) **Clinical review bias**: was the test measured independently of all the clinical information? **Observer/instrument variation**: test reproducibility	**Indeterminate test results**: percentage excluded because test was infeasible or result indeterminate	**Sample size**: sample size	**How were items chosen for inclusion on the scale?** Not reported. Authors state that checklist contains items relevant to assessing the quality and applicability of studies of diagnostic tests for inclusion in meta-analysis. Categorisation should be based on stated information in the study report or obtained from authors. Some commonly used criteria are not included because they are relevant to the quality of an individual article but not to that article's contribution to a meta-analysis. An example is whether the authors have estimated measures of test accuracy correctly and estimated confidence intervals **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Cooper, 1988[30] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: methodological review; authors' own tool		**Appropriate reference standard**: presentation of a reference standard: (a) tissue diagnosis, (b) other imaging procedure, (c) none	**Review bias (test, diagnostic and clinical)**: random order of imaging tests when comparing MRI with another procedure (done/not done). Blinding of investigator with regard to clinical history or other test results (done/not done) **Observer/instrument variation**: measurement of interobserver/ instrument variation in reading images (done/not done)	**Appropriate results**: appropriate use of the terms sensitivity, specificity, PPV or NPV, false positive or false negative and accuracy: (a) 3 or more terms used, (b) 1 or 2 terms used, (c) none. Appropriate presentation of data described by each term: (a) 2 or more terms, (b) 1 term only, (c) none. Appropriate calculation of the values for each of the described terms (done/not done). Presentation of quantitative data (complete/partial/none). Appropriate statistical analysis of quantitative data: (a) distribution curves with statistical analyses and/or ROC curves for each diagnosis, (c) none **Data table**: (b) qualitative grouping as in 2 × 2 table	**Protocol**: evidence of research planning: (a) prior protocol, (b) Institutional Review Board approval only, (c) neither	**How were items chosen for inclusion on the scale?** Not reported. Authors state that they selected 10 criteria that are considered important in assessing the precision or accuracy of the procedure as a clinical diagnostic measure **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: inter-reader scoring differences occurred in 14% of 540 observations **Topic area**: MRI

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Deeks, 1999[216] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: inappropriate sample selection – important to ensure that the distribution of disease severity and alternative diagnoses are representative of those prevalent in the setting where the test will be used	**Appropriate reference standard**: *Unsuitability of reference standard*: important to assess whether the reference standard is clearly defined when evaluating an article, whether it is the best test available and the likelihood that it could be wrong **Verification bias**: *Completeness of assessment of the reference standard*: important to assess whether sequential test selection processes may lead to incomplete assessments of a selected group of patients **Incorporation bias**: where index test is included in a battery of tests used to establish the reference standard diagnosis	**Review bias (test and diagnostic)**: non-independence of the experimental test and the reference standard; test should be undertaken blind to the other's results and according to a standard process	**Precision of results**: consider statistical significance and clinical significance of results		**How were items chosen for inclusion on the scale?** Not reported. References other scales and so appears to have been based on these[26,35,36,45,48] **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Deeks, 2001[225] **Aim**: guidelines for interpreting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: clinical and demographic characteristics fully described, complete **Population recruitment**: consecutive or randomly selected sample, recruited as single cohort unclassified by disease state, recruited from clinical setting and point in referral process where test would be used, selection and referral processes fully described	**Appropriate reference standard**: diagnosis likely to be close to the truth **Test execution**: application of test described in detail **Reference standard execution**: methods and tests described in detail **Verification bias**: results based on same tests and information in all patients, reference standard diagnosis available for all patients **Normal defined**: positive and negative diagnoses clearly described; diagnosis likely to be close to the truth **Treatment paradox**: test undertaken before treatment commenced	**Review bias (test and diagnostic)**: blinding procedures used to prevent knowledge of result of experimental test influencing reference diagnosis, made before treatment commenced. Blinding procedures used to ensure that test is undertaken without knowledge of reference diagnosis	**Indeterminate test results**: results reported for all patients, including those with 'grey zone' results		**How were items chosen for inclusion on the scale?** Not reported. Items grouped into 3 categories: sample of patients, reference diagnosis and experimental test **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Deeks, 2001[1] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: *Selection of study sample*: study population should be representative of the spectrum of diseases within which the test will be applied in practice *Aspects of the quality of reporting*: study should include clear descriptions of demographic characteristics and co-morbidities **Population recruitment**: *Selection of study sample*: ideal sample is a consecutive (or randomly selected) group of patients recruited from a clinically relevant population. Case–control studies subject to bias. Study samples should be selected from similar healthcare settings *Aspects of the quality of reporting*: study should include clear descriptions of source and referral history of patients	**Appropriate reference standard**: *Ascertainment of reference diagnosis*: selection of a good reference standard is crucial **Test execution**: *Aspects of the quality of reporting*: study should include clear descriptions of the experimental test **Description of reference standard execution**: *Aspects of the quality of reporting*: study should include clear descriptions of the reference standard **Normal defined**: *Aspects of the quality of reporting*: study should include clear descriptions of the positive and negative outcomes **Verification bias**: *Ascertainment of reference diagnosis*: problem when the decision to undertake the reference investigation is influenced by the result of the experimental test or other factors which indicate that this disease is unlikely **Incorporation bias**: where the result of the experimental test contributes to establishment of the reference standard **Treatment paradox**: *Other aspects of design*: both diagnostic tests should be undertaken before treatment is started, otherwise could be treated based on results of first test and then cured before second test and so misclassified	**Review bias (test and diagnostic)**: *Blinding*: each test should be undertaken and interpreted without knowledge of the result of the other	**Indeterminate test results**: *Other aspects of design*: important to include test results of all participants in the analysis		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Deyo, 1994[138] **Aim**: guidelines for interpreting study **Type of scale**: quality scores (method of scoring not reported) **Source of tool**: original tool	**Spectrum composition**: *Study population*: did the patient sample include an appropriate spectrum of mild and severe, treated and untreated disease, as well as patients with different but easily confused conditions? **Population recruitment**: was study setting and referral pattern for study subjects described?	**Appropriate reference standard**: *Test comparisons*: was the test independently compared with an appropriate reference standard? **Test execution**: *Test comparisons*: was the test described in a reproducible manner? *Strategies for use*: if the test is part of a group or sequence of tests, was its contribution to the overall validity of the sequence determined? **Incorporation bias**: *Test comparisons*: was the 'final diagnosis' established without the results of the test being evaluated?	**Review bias (test, diagnostic and clinical)**: *Test comparisons*: was the test assessed blindly relative to the reference standard or competing tests? **Clinical review bias**: was the test assessed in absence of clinical information? **Observer/instrument variation**: *Test comparisons*: was the reproducibility of the test results determined (intra- and interobserver/ instrument variation)?	**Appropriate results**: *Test comparisons*: were sensitivity and specificity calculated for comparing the test with the final reference standard diagnosis? **Utility of test**: *Strategies for use*: was the utility of the test determined, in terms of actual effects on patient care or outcomes?		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Dunn, 1995[140] **Aim**: guidelines for interpreting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: was the population from which this sample is drawn similar to the one in which the diagnostic test is going to be used? Has the diagnostic test been evaluated in a sample that included patients with the appropriate range of severity of symptoms, treated and untreated patients, and those with confused disorders? **Population recruitment**: was the sample drawn at random or through the use of subjective procedures?	**Appropriate reference standard**: has there been an independent 'blind' comparison of the results of the test with a reference standard of diagnosis? **Test execution**: if you wished to replicate the evaluation of the test yourself, have the procedures been described in sufficient detail to permit an exact replication?	**Review bias (test and diagnostic)**: has there been an independent blind comparison of the results of the test with a reference standard of diagnosis? **Observer/instrument variation**: has the diagnostic test been independently evaluated by different investigators under a wide range of conditions?	**Appropriate results**: what are the reported estimates of the test's sensitivity and specificity? **Precision of results**: do the authors provide the standard errors of these estimates (or, equivalently, do they provide corresponding confidence intervals?) Are these small or unacceptably large? **Utility of test**: if you decide to use a particular test or battery of tests, what are the benefits of being able to make a correct diagnosis? If a correct diagnosis makes no difference to prognosis or decisions concerning treatment, then why bother? What are the potential costs of making mistakes?	**Sample size**: was the sample size adequate for the job in hand?	**How were items chosen for inclusion on the scale?** not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Freedman, 1987[4] **Aim**: guidelines for reporting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: *Selection of patients*: an effort should be made to include a wide enough spectrum of patients to cover the composition of the target population. The imaging technique should be performed and reported in a routine hospital service environment Patients selected for study should be as nearly representative of the target population as possible	**Appropriate reference standard**: *Final diagnosis*: the reference standard should ideally be based on autopsy or histological evidence, particularly in oncology **Test execution**: *Performance of imaging technique*: to assist in diagnosis the radiologist should be presented with clinical information in accordance with normal clinical practice **Verification bias**: if techniques are sometimes missed the investigators should consider the possible effect on their results *Comparative studies*: no imaging technique should be allowed to influence the performance of a competing technique **Incorporation bias**: the imaging test under evaluation should not influence the final diagnosis, directly or indirectly **Normal defined**: if there is a problem of interpretation which is partly or wholly caused by a post hoc choice of cut-off point to separate negative from positive results, ROC curve analysis offers a possible solution	**Review bias (test)**: *Final diagnosis*: the imaging technique under evaluation should not influence the final diagnosis, directly or indirectly, i.e. the results of the imaging techniques should not be given to the clinician before diagnosis has been established. No imaging technique should be allowed to influence the performance of a competing technique **Clinical review bias**: a 2-stage reporting process should be used, first without the benefit of any clinical data and second taking those into account **Observer/instrument variation**: *Measures of diagnostic accuracy*: the reliability of the technique concerns the reproducibility of a test, e.g. from observer to observer or from test to test	**Appropriate results**: *Measures of diagnostic accuracy*: appropriate statistics should be presented, e.g. sensitivity, specificity, ROC curve, reliability, accuracy, PPV and NPV **Indeterminate test results**: *Excluding 'uninterpretable' test results*: it is important to report clearly the definition and the number of uninterpretable results for imaging techniques	**Sample size**: *Number of patients*: numbers of patients should be chosen in relation to the standard errors required for estimating sensitivity or specificity	**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: imaging

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Gifford, 1999[219] **Aim**: guidelines for interpreting study **Type of scale**: checklist **Source of tool**: original tool (NB for screening tests)	**Spectrum composition**: spectrum bias	**Verification bias**: work-up bias	**Review bias (test and diagnostic)**: review biases **Observer/instrument variation**: *Inter-rater and intrarater reliability*: kappa values for a given test should exceed 0.5 or 0.7 for inter-rater evaluations and 0.9 for intrarater evaluations *Internal consistency*: Cronbach's coefficient alpha often used to measure this, score > 0.8 is considered excellent, > 0.7 good, and < 0.4 poor *Content validity*: whether it looks as though the questionnaire measures what the test is designed to measured	**Appropriate results**: *Construct validity*: when a new test relates well to other measures as hypothesised, usually a reference standard test. A kappa statistic is often used to calculate agreement between 2 different tests **Criterion validity**: predictive validity (whether a new test predicts future performance); concurrent validity: sensitivity, specificity, area under ROC curve		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general/dementia

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Greenhalgh, 1997[48] **Aim**: guidelines for interpreting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: did this validation study include an appropriate spectrum of patients? The test should be verified on a population which includes mild and severe disease, treated and untreated subjects, and those with different but commonly confused conditions	**Appropriate reference standard**: has the test been compared with a true reference standard? **Verification bias**: has work-up bias been avoided? Did everyone who had the new diagnostic test also have the reference standard, and vice versa **Normal defined**: has a sensible 'normal range' been derived?	**Review bias (test and diagnostic)**: has expectation bias been avoided? All comparisons of diagnostic studies with a reference standard should be blind **Observer/instrument variation**: was the test shown to be reproducible? It is important to confirm that reproducibility of one observer and between different observers is at an acceptable level	**Precision of results**: were confidence intervals given? **Utility of test**: is this test potentially relevant to my practice? Does the test help me? Does it identify a treatable disorder? Would I use it in preference to the test I use now? Could I afford it? Would it change the probabilities for competing diagnoses sufficiently for me to alter my treatment plan? What are the features of the test as derived from this validation study? The sensitivity, specificity and other crucial features of the test should be at an acceptable level for the condition being studied. Has this test been placed in the context of other potential tests in the diagnostic sequence?		**How were items chosen for inclusion on the scale?** Not reported. Author states that she drew on a number of different sources[6,26,35,36,217] **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area:** general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Greiner, 2000[212] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: *Selection of reference populations*: the reference population is sufficiently described (time, location, animal characteristics such as breed, age, gender, etc.). The reference population should reflect the target population and include an appropriate spectrum of disease and spectrum of other conditions **Inclusion criteria**: *Selection of reference populations*: selection criteria must be stated and should reflect the testing situation *Sampling of the reference population*: exclusion or inclusion criteria are described **Population recruitment**: *Selection of reference populations*: the sampling frame should be an unbiased representation of the reference population; the sampling procedure should be described in detail. Random and systematic sampling are the preferred options	**Appropriate reference standard**: *Reference standard*: the choice of the reference method is justified (being more accurate than the new is a necessary condition) *Discussion of results*: if the reference standard is imperfect this should be discussed in relation to the effect on the study results **Test execution**: *Performance of test and reference standard*: the testing protocols are sufficiently described **Description of reference standard execution**: *Reference standard*: method is fully described or referenced **Normal defined**: *Performance of test and reference standard*: definition of negative and positive results	**Review bias (test and diagnostic)**: *Performance of test and reference standard*: results of test and reference standard are interpreted blindly	**Appropriate results**: *Presentation of results*: parameter estimators are explained by formulae. Sensitivity and specificity are always required; additional parameters may be presented as necessary. ROC analysis should be presented for test outcomes measures on ordinal, interval or ratio scales **Precision of results**: *Presentation of results*: estimates are presented together with sample sizes and confidence intervals **Indeterminate test results**: *Presentation of results*: the number of uninterpretable and intermediate results and reasons for missing data are given **Data table**: *Presentation of results*: the source 2×2 table should be displayed **Utility of test**: *Discussion of results*: the test performance parameters should be discussed in relation to the study design and the intended or current use of the test	**Sample size**: *Sampling of the reference population*: sample sizes must be stated and should reflect the degree of the required statistical certainty **Objective**: the test purpose and the analytical unit are described **Protocol**: the test protocol is sufficiently described	**How were items chosen for inclusion on the scale?** Adapted from Mulrow (1989)[45] and Jaeschke (1994).[35,36] Authors state that the checklist is an excerpt of existing guidelines supplemented with their own epidemiological comments **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** See details above **Level of inter-rater reliability**: not reported **Topic area**: veterinary medicine

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Guyatt, 1992[143] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: systematic review; authors' own tool	**Population recruitment**: consecutive recruitment, patient consent for invasive procedure, explicit definition of anaemia	**Test execution**: *Interventions*: specified method of testing (i.e. how laboratory tests were carried out)	**Review bias**: *Outcome measures*: bone marrow examined by 2 or more readers blinded to results of other tests			**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: populations: absolute agreement = 0.72, kappa = 0.40; interventions: absolute agreement = 0.86, kappa = 0.49; outcome: absolute agreement = 0.84, kappa = 0.63 **Topic area**: diagnosis of anaemia
Haynes, 1995[224] **Aim**: guidelines for reporting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: clearly identified comparison groups, at least one of which is free of the target disorder	**Appropriate reference standard**: either an objective diagnostic standard or a contemporary clinical diagnostic standard with demonstrably reproducible criteria for any subjectively interpreted component	**Review bias (test and diagnostic)**: interpretation of the test without knowledge of the diagnostic standard result; interpretation of the diagnostic standard without knowledge of the test result			**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Heffner, 1998[213] **Aim**: guidelines for reporting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: *Spectrum bias*: assessment studies need to simulate the reality of clinical practice where diagnostic tests are used to discriminate between the presence and absence of disease in patients who present with suggestive manifestations and who could reasonably have the target disorder. Assessment studies must carefully describe their sample population in terms of demographic and clinical features **Population recruitment**: reassurances should be stated that patients were enrolled consecutively or through a randomisation process. The referral patterns for examination of the patients also need to be described to determine the magnitude and expected direction of the selection bias	**Appropriate reference standard**: *Reference standard*: studies must clearly define the reference standard, and should select reference standards that are the most definitive method for assessing the presence or absence of disease considering the relative accuracy and feasibility of the reference standard tests that are available **Change in technology of test**: *Temporal changes*: it would be expected that diagnostic accuracy of tests would change over time as better reference standards are developed **Description of reference standard execution**: *Reference standard*: studies must clearly define how the reference standard is applied **Verification bias**: occurs if the diagnoses for patients with different results on the new diagnostic test are not equally likely to be confirmed or verified by an existing reference standard test **Normal defined**: *Measures of test efficacy*: diagnostic tests with continuous or ordinal values usually undergo dichotomisation; this simplifies test interpretation but makes the test's results dependent on the choice of cut-off value *Determination of cut-off points*: cut-off point may be selected according to the trade-off between sensitivity and specificity appropriate to the target condition	**Review bias (test and diagnostic)**: *Test review bias*: it is important to interpret the test result and the reference standard independently from each other **Observer/instrument variability**: assessment studies need to indicate how reproducible the results of the new test are and the standards that ensure a high level of reproducibility	**Appropriate results**: *Sensitivity, specificity, predictive values, LRs, odds ratios, ROC curves*: these are most commonly used measures of the discriminative properties of diagnostic tests. Author gives details of how to calculate these **Precision of results**: *Precision estimates of diagnostic accuracy measures*: precision can be described by calculating confidence intervals **Indeterminate test results**: *Uninterpretable test results and reproducibility*: uninterpretable results may be excluded from analysis if the test is repeatable and the cause of the uninterpretable result is random. If the cause is not random investigators should report the proportion of studies that was uninterpretable and describe the potential effects on the calculated discriminative properties of the test **Analysis of subgroups**: *Standard of subgroup analysis*: it is important for investigators to assess the discriminative properties of new diagnostic tests in pertinent subgroups of their patient population to allow their findings to be generalised for other patient population		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: pleural cavity

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Heffner, 1998[199] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: methodological review; authors' own tool	**Spectrum composition**: *Spectrum composition*: had to describe the case mix of the study population adequately to allow conclusions to be drawn regarding the diagnostic accuracy of the tests and their generalisability to other patient populations. Study standard considered fulfilled if article provided 3 of the 4 following descriptions: age distribution, gender distribution, summary of presenting clinical symptoms and/or disease stage, and inclusion criteria for study subjects *Indicated study sample*: evaluated patients appeared to have a realistic likelihood that they had the disorder undergoing diagnostic evaluation **Population recruitment**: *Patient sampling techniques*: study had to include consecutive patients or a random sample of consecutive patients in order to avoid sampling bias	**Appropriate reference standard**: *Clear definition of reference standard/suitability of reference standard*: reference standard had to be best available method for assessing the presence or absence of disease considering the relative accuracy and feasibility of alternative methods **Verification bias**: *Work-up (or verification) bias*: met for cohort studies if all of the study patients were submitted to both the diagnostic test under evaluation and the reference standard procedure. The standard could also be fulfilled with suitable follow-up if the reference standard was unfeasible (too expensive or invasive) for patients with negative test results. For case–control studies wherein the diagnostic test followed the reference standard procedure the standard was met if the test results were stratified by the clinical factors that prompted performance of the reference standard procedure	**Review bias (test and diagnostic)**: *Test review bias*: for cohort studies wherein the reference standard procedure always followed the diagnostic test, a statement was required that the reference standard procedure was interpreted independently of the diagnostic test. For cohort studies wherein the reference standard procedure was sometimes performed before and at other times after the diagnostic test, a statement was required that both the diagnostic test and the reference standard procedure was interpreted independently. For case–control studies wherein the reference standard preceded the diagnostic test, a statement was required indicating that the diagnostic test was interpreted without knowledge of the reference standard result **Observer/instrument variation**: *Test reproducibility*: for tests that depended on observer interpretation at least some of the test subjects needed	**Appropriate results**: *Head-to-head comparisons*: for studies that compared 2 or more diagnostic tests with a reference standard, these had to be compared using appropriate statistical comparisons *Summary measures of diagnostic accuracy*: articles were reviewed to determine whether some or all of the following measures were reported: sensitivity and specificity, predictive values, LRs, odds ratios and values for the ROC curves. **Precision of results**: *Precision of summary measures*: accepted measures of precision included confidence intervals and SEs **Indeterminate test results**: *Indeterminate test results*: 2 standards; the first required a statement regarding the existence and frequency of indeterminate results generated in the study; the second required that the study indicated whether indeterminate results were included or excluded in calculations of test accuracy **Analysis of subgroups**: *Analysis of pertinent subgroups*: article had to evaluate the diagnostic accuracy of the tests in one or more explicitly defined subgroups		**How were items chosen for inclusion on the scale?** The methodological standards were compiled from established resources for assessing the quality of investigational design for studies in diagnostic test research.[5,28,30,35,39,66,130,152,206,246–248] Additional standards assessed the descriptors used to quantitate the diagnostic accuracy of tests, and the degree to which the evaluated tests' clinical value in patient care was discussed **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: kappa values for the assessed standard between the two primary rates ranged from 0.72 to 0.92 **Topic area**: pulmonary medicine

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
		Head-to-head comparisons: for studies that compared 2 or more diagnostic tests with a reference standard all study patients or a random sample of study patients had to be subjected to all of the compared diagnostic tests **Normal defined**: *Method for determining the diagnostic threshold*: method used for selecting decision thresholds had to be clearly described; method had to conform to accepted techniques; included ROC analysis and selection of decision thresholds that fulfilled predefined diagnostic goals	to have been evaluated for a summary measure of observer/instrument variation. For tests performed without observer interpretation a summary measure of instrument variability needed to be provided	**Utility of test**: *Test acceptability*: article had to make a statement regarding the acceptability of the new test in terms of factors such as minimum detection levels and cross-reactivity of biochemical tests, degree of measurement errors, required personnel and equipment, cost, acceptability to patients, dose and pharmacokinetics of any drugs or agents *Incremental value*: the relative value of the new test had to be compared with existing tools and the incremental value be discussed		

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Hoffman, 1991[133] **Aim**: assess study quality **Type of scale**: levels of evidence **Source of tool**: systematic review; authors' own tool	**Spectrum composition**: *Cohort assembly*: high: wide spectrum of clinical severity from clinical practice-based referrals; intermediate: retrospective, index test selection bias, reference standard selection bias, limited spectrum with referral filtering; low: work-up bias or cohort assembly not described *Adequacy of clinical description*: high: good clinical description with complete demographics; intermediate: sketchy clinical description; low: no clinical description other than low back pain **Population recruitment**: *Cohort assembly*: high: prospective; intermediate: retrospective	**Appropriate reference standard**: *Technical quality of reference standard*: high: high-resolution CT scan, MRI, water-soluble myelogram, surgical findings, overall clinical impression after diagnostic imaging; intermediate: low-resolution or unspecified CT scan, MRI, oil-based or unspecified myelogram; low: unacceptable or unspecified reference standard **Verification bias**: *Uniform application of reference standard*: high: single reference standard applied to all analysed cases; intermediate: mixed reference standards, all cases analysed; low: no acceptable reference standard applied to any case	**Review bias (test and diagnostic)**: *Independence of interpretation*: high: no independence problems exist; intermediate: only one of two biases (test review or diagnosis review) present or unable to be excluded; low: test review and diagnosis review biases present, or no information to assess independence		**Sample size**: high ≥ 35 diseased and ≥ 35 non-diseased; intermediate ≥ 35 diseased and < 35 non-diseased; low: < 35 diseased	**How were items chosen for inclusion on the scale?** Based on previously proposed assessment categories and quality criteria.[5,6,45,66] **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: the overall concordance for category rating was 0.74 (kappa = 0.61), the kappa values for the individual ratings were: reference standard technical quality, 0.33; reference standard application, 0.65; independence, 0.68; clinical description, 0.26; cohort assembly, 0.34; sample size, 1.0 **Topic area**: thermography for lumbar radiculopathy

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Holleman, 1995[127] **Aim**: assess study quality **Type of scale**: levels of evidence **Source of tool**: review (not systematic); authors' own tool	**Population recruitment**: Grade A: independent, blind comparison of sign or symptom with a reference standard of diagnosis among a large number of consecutive patients suspected of having the target condition Grade B: independent, blind comparison of sign or symptom with a reference standard of diagnosis among a small number of consecutive patients suspected of having the target condition Grade C: independent, blind comparison of sign or symptom with a reference standard of diagnosis among non-consecutive patients suspected of having the target condition or non-independent comparison of sign or symptom with a reference standard of diagnosis among sample of patients who obviously have the target condition plus, perhaps, normal individuals or non-independent comparison of sign or symptom with a standard of uncertain validity	**Appropriate reference standard**: see standards under Population recruitment **Verification bias**: see standards under Population recruitment	**Review bias**: see standards under Population recruitment			**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported, assume not **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Irwig, 1994[130] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: Methodological review; modified tool		**Appropriate reference standard**: the reference standard must be clearly defined and should be the best available method of assessing the presence or absence of the disease of interest **Verification bias**: diagnostic accuracy should be assessed in consecutive patients who present with the clinical problem of interest. Verification bias occurs when the reference standard has been assessed on patients sampled differentially in the categories of test results	**Review bias (test and diagnostic)**: *Independence of observations*: those involved in assessing test results should be blind to the results of the reference standard. Likewise, assessors of the reference standard should be blind to the test result			**How were items chosen for inclusion on the scale?** Not reported; based on Begg (1987)[61] **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general
Jaeschke, 1994[35] **Aim**: guidelines for interpreting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice? The true pragmatic value of a test is established only in a study that closely resembles clinical practice	**Appropriate reference standard**: was there an independent blind comparison with a reference standard? Must assure yourself that an appropriate reference standard has been applied to every patient along with the test under investigation **Test execution**: were the methods for performing the test described in sufficient detail to permit replication? The description should cover all uses that are important in the preparation of the patient, the performance of the test, and the analysis and interpretation of its results **Verification bias**: did the results of the test being evaluated influence the decision to perform the reference standard?	**Review bias**: was there an independent blind comparison with a reference standard? Have to assess whether the reference standard and test results were assessed independently of each other **Observer/instrument variability**: will the reproducibility of the test result and its interpretation be satisfactory in your setting? Ideally, the reproducibility of the test results should be reported	**Appropriate results**: are LRs for the test results presented or data necessary for their calculation included? **Utility of the test**: will patients be better off as a result of the test? Will the results change patient management?		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Jensen, 1999[211] **Aim**: assess study quality **Type of scale**: levels of evidence **Checklist**: quality score **Source of tool**: original tool	**Spectrum composition**: *Failures in the planning of the study*: the study population is inadequate	**Appropriate reference standard**: *Deficiencies in the objectives of the study*: The reference standard contained arbitrary elements, e.g. the dependence of the validity on the different elements is not known or investigated *Failures in the planning of the study*: the choice of reference standard is inadequate **Test execution**: *Deficiencies in the objectives of the study*: the definition of the diagnostic test is inadequate for the requirements *Faults in the presentation of results*: the study is not reproducible **Verification bias**: *Failures in the planning of the study*: there is the possibility of verification bias **Normal defined**: *Failures in the planning of the study*: in binary test calculated from quantitative parameters the criteria for positive test results are not described or are not clear	**Review bias**: *Failures in the planning of the study*: reciprocal blinding is not assured	**Precision of results**: *Faults in the presentation of results*: confidence intervals are not presented **Test utility**: a diagnostic test was evaluated which cannot provide any use of information gain **Indeterminate test results**: uninterpretable results are not mentioned or are ignored **Analysis of subgroups**: the correlation between sensitivity and specificity and covariables was not investigated	**Sample size**: a sample size calculation was not performed or the sample size was too small **Objectives**: *Deficiencies in the objectives of the study*: a phase of the test evaluation was conducted before the information from earlier phases was available. A test, which is proposed as an alternative for an existing standard test, is evaluated in isolation and not in comparison of both tests on the same sample. A test that is part of a diagnostic strategy is evaluated in isolation. In tests that will form part of a diagnostic strategy the aspect of the declining specificity is neglected **Protocol**: *Failure in the planning of the study*: there is no written study protocol	**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Kelly, forthcoming[205] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: *Patient cohort bias*: are the study groups' clinical details described? Are the study groups' pathological details described? Are the study groups' co-morbid details described? **Inclusion criteria**: *Patient filtering bias*: are specific inclusion criteria stated for those included/excluded? Is cointervention bias (treatment paradox) present or avoided via the inclusion criteria? **Population recruitment**: *Referral bias*: is the establishment where the study was undertaken stated? Is the establishment from where the patients were referred stated? Is access to the establishment described, i.e. open access, referral based, public or private, etc.?	**Disease progression bias**: is disease progression bias present for the test under evaluation? **Verification bias**: *Biases associated with application of the reference standard*: is verification bias present? Is work-up bias present? Is incorporation bias present? **Treatment paradox**: is cointervention bias present?	**Review bias**: *Independence of interpretation biases*: is diagnostic review bias present? Is test review bias present? Is comparator review bias present? Is clinical review bias present? **Observer/instrument variation**: is there a single observer of the diagnostic test under evaluation? If no, are results reported separately for each observer? Is any attempt made to assess Interobserver/instrument variation? Are the diagnostic test results taken from a consensus decision? Is any attempt made to assess intraobserver/instrument variation?	**Indeterminate test results**: *Withdrawal bias*: are results reported for all patients who received verification? Are there any indeterminate test results? Are there any patients lost to follow-up?		**How were items chosen for inclusion on the scale?** Not reported. Authors state that the checklist was designed to assess individual study quality by containing specific questions applicable to each of the potential biases, while maintaining a broad applicability over all diagnostic performance studies. In order to be able to answer the questions in a reproducible, objective manner, very specific guidelines are required that may require slight modification between diagnostic specialities **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: medical imaging

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Kent, 1992[131] **Aim**: assess study quality **Type of scale**: levels of evidence **Source of tool**: original tool	**Spectrum composition**: High quality (grade A): spectrum of patients studied should be similar to that encountered in practice and not filtered to favour only patients with severe disease Good quality (grade B): sample a more limited spectrum of patients, typically reflecting the referral bias of university centres with more severely ill patients	**Appropriate reference standard**: *High quality (grade A)*: reference standards are required to permit independent assessment of disease presence, extent or functional severity without the threat of biases *Weak or non-contributory studies*: lack of an independent reference or reference standards **Verification bias**: *Weak or non-contributory studies*: where positive results may have been favoured by using results to define cases **Incorporation bias**: *Weak or non-contributory studies*: studies which declare the final diagnosis using the test results under study	**Review bias (test)**: *Good-quality studies (grade B)*: free of other procedural flaws that promote interaction between test result and disease determination *Weak or non-contributory studies*: involve interpretation of test results while knowing the ultimate diagnosis, or by declaring the final diagnosis using the test results under study	**Indeterminate test results**: high-quality (grades A and B): should include and acknowledge indeterminate results in analyses	**Sample size**: high quality (grade A): should have sufficient sample size (> 35); good quality (grade B): should sample > 35 studies	**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: MRI

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Kent, 1992[151] **Aim**: assess study quality **Type of scale**: levels of evidence **Source of tool**: systematic review; authors' own tool	**Spectrum composition**: *Clinical description*: high: description includes demographics (age, gender), duration of symptoms, and percentage of cases with usual symptoms and physical findings; intermediate: incomplete demographic or clinical description; low: no description other than 'low back pain' or 'radiculopathy' **Population recruitment**: *Cohort assembly*: high: wide spectrum of clinical severity, enrolled prospectively from typical practice sources; intermediate: retrospective case finding, limited spectrum of types of severity, referral filtering due to enrolments after tests ordered from specialised centres; low: work-up bias present or procedure for assembling cohort not described	**Appropriate reference standard**: *Reference standard quality*: high: surgical findings or overall impression after technical imaging and follow-up, explicit criteria for surgical or clinical final diagnosis; intermediate: surgical or clinical follow-up without explicit criteria; low: unacceptable or unspecified reference standard **Test execution**: *Index test technical quality*: high: best state-of the art techniques; intermediate: average techniques found in usual practice; low: obsolete, technically flawed or not described **Verification bias**: *Application of reference standard*: high: single reference standard applied to all analysed cases; intermediate: mixed reference standards, all cases analysed; low: acceptable reference standard not applied to all cases	Review bias (test and diagnostic): *Independence of interpretations*: high: study protocol prevented both test review and diagnosis review biases; intermediate: one of two biases present or cannot be excluded; low: both biases present, or no information to assure independence of test results and reference standard determinations		**Sample size**: high: ≥ 35 diseased and ≥ 35 non-diseased; intermediate: ≥ 35 diseased and < 35 non-diseased or reverse; low: < 35 diseased and < 35 non-diseased	**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: diagnosis of lumbar spinal stenosis using CT, MR and myelography

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Khan, 2001[209] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: original tool	**Population recruitment**: ideal: consecutive recruitment of eligible patients; non-ideal: non-consecutive recruitment or convenience sampling	**Appropriate reference standard**: ideal: measurement parameter and cut-off accepted standard; non-ideal: partially reported **Verification bias**: *Follow-up*: ideal: all those having diagnostic test had reference standard; non-ideal: application of reference standard based on test result **Normal defined**: *Diagnostic test*: ideal: measurement parameter and cut-off reported; non-ideal: partially reported	**Review bias (test and diagnostic)**: *Blinding*: ideal: blinding between diagnostic test and reference standard; non-ideal: not blind			**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Kobberling, 1990[40] **Aim**: guidelines for interpreting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: *Structure description*: must be carefully described and should include details of accompanying diseases (in particular those that could have a systematic influence on the test result), additional characteristics that may have an influence on the test result (e.g. age, gender, weight, alcohol, tobacco, drugs, social status) **Inclusion criteria**: *Inclusion and exclusion criteria*: must be defined for the patients that are recruited to the study. The main inclusion criterion should be that for the particular patient the application of the diagnostic test can occur in an actual clinical situation. If this inclusion criterion is fulfilled then only serious reservations should prevent the patients from taking part in the study. If for financial or capacity reasons only a part of the patients who fulfil the inclusion and exclusion criteria can be included in the study, then this should occur independently of other characteristics, e.g. randomised selection **Disease prevalence/severity**: *Structure description*: should include the prevalence of the disease of interest in the collective investigated	**Verification bias**: *Establishment of the diagnosis*: it must be guaranteed that patients with positive and negative results are subjected to the same diagnostic procedures. This regimen must be established before the study has begun. The result of the test should not affect the subsequent diagnosis. If it is not justifiable to subject all test-negative or test-positive patients to the measures of a standardised diagnosis then a reduction in the diagnostic methods can be undertaken for a part of the patients. To avoid a systematic error through the selection (so-called selection or work-up bias) the allocation of study participants to the diagnostic methodology should not be based on selection mechanisms but should be randomised. Only in certain cases can careful observation of the patient (follow-up) replace the actual establishment of a diagnosis	**Review bias (test and clinical)**: *Observation methods*: those who diagnose the disease status should remain blind to the test result. Similarly, those involved in the assessment of the diagnostic test should have no information on the disease status of the respective patients. Other clinical information on the patients (e.g. gender) can be made available for the evaluation of the test so long as this is necessary for a meaningful judgement	**Appropriate results**: *Establishment of the diagnosis*: the reliability of the diagnosis should be provided in as much as this is possible **Data table**: *Evaluation*: the results (numbers) should be presented in the form of a contingency table so that the calculations of the test parameters can be performed. The restriction of the diagnostic procedures to a part of a test-positive and/or test-negative group should be taken into consideration so as not to distort the evaluation	**Protocol**: a study protocol should be initially formulated in which the study design is described Patients must give their consent before being recruited	**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Kraemer, 1992[218] **Aim**: guidelines for interpreting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: in what clinical population is the test proposed for use? **Population recruitment**: will sampling be naturalistic, retrospective or prospective? How can the sampling be done so as to obtain a representative sample from that population?	**Appropriate reference standard**: if the test is to be evaluated as a prognostic test, will the follow-up be fixed or variable? If the follow-up is to be fixed, fixed at what time? The follow-up time should be long enough to give the test a fair chance and consistent with the specific medical purpose one has in mind. Is there any internal or external standard provided for the performance of an excellent test? **Test execution**: what is the disorder? What diagnosis is to be used? Is the diagnosis clinically valid and reliable? A test should be proposed with a specific and well-defined purpose, and should only be used in practice if the outcome can be specifically interpreted. What are the tests under evaluation, their protocols, responses and references? What are the test costs? Is this a single test or is this to be considered a battery of tests? What is the quality of each single test under evaluation and for each population under evaluation? If there is a battery of tests under consideration, has the battery been appropriately evaluated? **Normal defined**: has the optimal referent for a test, or the optimal first test in a battery test been appropriately selected?	**Review bias**: how are the blinding of the diagnosis and test results assured? If there are multiple tests in a battery under evaluation, are these blinded to each other?	**Appropriate results**: was the database properly compiled and thoroughly checked for errors? For each single test under evaluation, for each test in the battery under evaluation, and for each population under evaluation, have the descriptive statistics been properly computed? **Indeterminate test results**: how are dropouts and missing response avoided? **Dropouts**: how are dropouts and missing response avoided? **Utility of test**: what are the clinical benefits in this situation?	**Sample size**: how large a sample size is needed?	**How were items chose for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Lang, 1997[223] **Aim**: guidelines for reporting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: specify the stage of the condition for which the test is appropriate: pathological component of severity or extent of disease, clinical component of the severity of chronicity of symptoms and a co-morbid component of other diseases, not directly related to the disease under study, that may affect test results	**Appropriate reference standard**: the accuracy of the reference standard should be addressed **Test execution**: identify the purpose of the test: the medical condition or diagnosis that the test is intended to define and the population for which the test is appropriate should be identified. Describe the biological principle on which the test is based: knowing how the test works helps readers to evaluate the validity of the test. Describe how the test is to be administered **Verification bias**: the index test result must be independent of the verification of disease (work-up bias) **Normal defined**: explain the meaning or clinical meaning of a positive test result; could give a diagnostic definition (the range of measurements over which the condition is absent and beyond which the condition is likely to be present) or a therapeutic definition (range of measurements over which a therapy is not indicated). Give the rationale for selecting a given cut-point	**Review bias**: the index test result must be independent of the verification of disease; diagnostic review bias, incorporation bias **Observer/instrument variation**: report differences in how the test was administered; interobserver reliability, intraobserver reliability. Report differences in how the test sample was processed. Report differences in the conditions under which the patient was tested. Report validity of the test under study and the reference standard to which it was validated. Report reliability of the test: can be affected by differences in how the test was administered, test sample was processed, conditions under which patient was tested, and intra-/interobserver reliability	**Appropriate results**: report the positive and negative LRs of the test. When a diagnostic test is an essential part of the research, and when its interpretation depends on a cut-point on a continuum, illustrate its characteristics with an ROC curve. Report the PPV and NPV, as well as the prevalence of the disease associated with these values. When reporting the use of 2 or more diagnostic tests in combination, indicate the order in which the tests were conducted, the characteristics of each, and the contribution of each test to the final results **Precision of results**: report the diagnostic sensitivity and specificity of the test, including the associated confidence intervals **Indeterminate test results**: explain the meaning of equivocal results and how such results were incorporated into the calculation of the test's characteristics, including intermediate results, indeterminate results and uninterpretable results **Analysis of subgroups**: if appropriate, identify any subgroups for which the test may be particularly effective **Test utility**: describe the medical costs and benefits to society of adopting the test, including the impact on patients misdiagnosed or misclassified as a result of the test. Describe how the test compares with similar tests	**Sample size**: report the number and proportion of patients with and without the disease who were tested to determine the specificity and sensitivity	**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Lensing, 1993[200] **Aim**: assess study quality **Type of scale**: levels of evidence **Source of tool**: systematic review: modified tool	**Spectrum composition**: identification of patients selected for study, e.g. type of surgery **Population recruitment**: inclusion of consecutive patients	**Test execution**: description of the tactics for carrying out the test in sufficient detail to permit its exact replication **Normal defined**: establishment of a priori objective criteria for a normal and an abnormal leg scan	**Review bias (test, diagnostic and clinical)**: independent comparison with the reference standard for venous thrombosis (contrast venography) by investigators blinded to clinical and prior test information **Observer/instrument variation**: determination of the reproducibility of the test results and their interpretation			**How were items chosen for inclusion on the scale?** Not reported. Modification of Sackett (1985)[6] **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: DVT
Liddle, 1996[207] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: is the spectrum of patients with and without the disease appropriate for the proposed use of the test? What factors could affect the accuracy of the test, e.g. special diets? What are the characteristics of the population and study setting?	**Appropriate reference standard**: is the test compared with a valid reference standard? **Verification bias**: is the decision to perform the reference standard independent of the test results (avoidance of verification)?	**Review bias (test and diagnostic)**: are the test and reference standard measured independently (i.e. blind to the other test result)? If multiple tests are compared, are the tests assessed independently of each other on the same patient or are they performed on randomly allocated patients?	**Dropouts**: *Loss to follow-up*: what percentage of the described study group was not included in the analysis?		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general
Mant, 1995[217] **Aim**: guidelines for interpreting study quality **Type of scale**: checklist **Source of tool**: original tool		**Appropriate reference standard**: assess the reference standard		**Data table**: check the 4-box analysis	**Objectives**: identify the key clinical question	**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Mower, 1999[67] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: *Spectrum and subgroup bias*: does the clinical population (in whom the test will be used) differ from the study population? **Disease prevalence/severity**: *Context bias*: does test interpretation vary with changes in disease prevalence/severity?	**Appropriate reference standard**: *Absence of a definitive test*: was a definitive standard used in assessing the presence or absence of disease? **Change in technology of index test**: *Temporal changes*: can technical advances improve test interpretation or disease detection? **Verification bias**: *Work-up bias*: were patients included in the study on the basis of previous tests or diagnostic work-up? **Normal defined**: *Variations in positivity criteria*: can change in positivity criteria produce variation in test accuracy?	**Review bias (test, diagnostic and clinical)**: are determination of disease status and test results made independently? Investigators must be aware that if clinical information is provided it may enhance the apparent efficacy of the test **Incorporation bias**: where results of index test are actually used to establish the final diagnosis **Observer/instrument variation**: *Reproducibility*: is test reproducibility documented? *Temporal changes*: can operator experience improve test interpretation or disease detection?	**Indeterminate test results**: *Unclear results*: are intermediate, indeterminate and uninterpretable results adequately presented and not just ignored or incorporated into indices in a dichotomous manner? **Analysis of subgroups**: *Spectrum and subgroup bias*: does the test perform well in specific patient subgroups and are these groups well described?		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Mulrow, 1989[45] **Aim**: assess study quality **Type of scale**: quality score **Source of tool**: original tool	**Spectrum composition**: were individuals with and without disease included in the evaluation? Was the study population appropriate for evaluating the proposed use of the diagnostic test? Was a wide spectrum of severity of disease patients included in the case group? Were patients with a wide spectrum of co-morbid diseases included in the control group? Were patients with co-morbid conditions included in the case group? Was the source of the study population described? Were demographic and clinical characteristics of study patients described? **Inclusion criteria**: were the inclusion and exclusion criteria that were used to select study patients described?	**Appropriate reference standard**: was an appropriate reference standard used? **Test execution**: was the diagnostic test being evaluated appropriately performed in a standardised manner? Was the proposed use of the test described? **Description of reference standard execution**: was the reference standard appropriately performed in a standardised manner in all patients? **Normal defined**: was the normal test value adequately defined? **Reference defined**: was a normal reference standard adequately defined?	**Review bias (test and diagnostic)**: were the interpretations of the reference standard and of the diagnostic test applied independently? **Observer/ instrument variation**: was the precision (reproducibility) of the test described?	**Appropriate results**: were data presented in enough detail to calculate appropriate test characteristics? **Indeterminate test results**: were uninterpretable results enumerated and described?	**Sample size**: was an appropriate sample size considered?	**How were items chosen for inclusion on the scale?** *Panel members*: scale development was accomplished by 14 panel members; all had practical experience in using diagnostic tests and 9 had training in epidemiology *Identification and weighting of questions*: *Step 1*: panel of five members who had studied relevant literature met in three committee meetings and explicitly identified 16 questions that addressed the adequacy of a diagnostic test evaluation *Step 2*: a second independent panel of 9 members was asked to complete an open-ended questionnaire based on guidelines for diagnostic test evaluations; members were asked to comment on the relative importance of the 8 published McMaster criteria, whether these criteria warranted further clarification or expansion, and whether additional criteria were needed. 2 editors compiled answers from these questionnaires into 20 explicitly defined questions *Step 3*: all questions were combined into a single 28-item closed-ended questionnaire. This questionnaire included 8 questions originally specified by the interactive panel only, 8 questions specified by both panels and 12 questions specified by the independent panel *Step 4*: the 28-item questionnaire was sent to all 14 panel members. They were asked to assess whether the questions were appropriate and necessary to consider for reviewing the quality of a diagnostic test evaluation. Question assessments were scored as either paramount (absolutely essential criteria) or on a scale of 0–5 *Step 5*: scores for each question were collated and returned as feedback to the 14 panel members. They were asked to rescore each of the 28 questions after referring to their own previous answers and the previous answers given by other panel members. In addition, for questions scored as paramount, panel members were now asked to indicate their degree of certainty in assigning that ranking *Final scoring and editing*: scores for each of the 28 questions were averaged; to decide which questions should remain in the final

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
						questionnaire, 2 editors applied cut-points (determined a priori) to the average scores. Questions that had been ranked paramount without reservation by more than half of the panel members were considered paramount in the final questionnaire. Questions receiving an average rating of < 3 were not included. The remaining questions were retained along with their average scores **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** See above **Level of inter-rater reliability**: a test set of 16 articles was evaluated by 3 reviewers independently to rank the studies. Pearson correlation coefficients for the summary ranking of the articles ranged from 0.93 to 0.96. Kappa values of 1.0 were obtained for 11 of the 19 questions in the scale; the remaining 8 questions had kappa values that ranged from 0.66 to 0.85 **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Panzer, 1986[75] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: systematic review; authors' own tool	**Spectrum composition**: *Study population*: control patients included, spectrum of control patients appropriate, with controls selected who were suspected of having the disease of interest **Population recruitment**: *Study population*: source of patients described	**Appropriate reference standard**: was the reference standard an independent procedure, e.g. pathology, surgery? **Test execution**: recorded field strength of MR image and type of CT used **Execution of reference standard**: *Test performance and interpretation*: reference standard defined **Verification bias**: *Test performance and interpretation*: tests performed in random order	**Review bias (test, diagnostic and clinical)**: *Test performance and interpretation*: blinded interpretation, independent procedure without access to other imaging studies or biasing clinical information **Observer/instrument variation**: *Test performance and interpretation*: observer variation measured	**Appropriate results/ indeterminate test results**: *Analysis*: details of test results provided for all cases, sensitivity and specificity calculated, statistical analysis performed		**How were items chosen for inclusion on the scale?** Standards were derived from published articles and books addressing issues in the design of research evaluating the efficacy of new diagnostic tests **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: MRI and CT

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Philbrick, 1980[155] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: modified tool	**Spectrum composition**: adequate identification of the groups selected for study. Statement of eligibility criteria enough to allow the same group of patients to be selected if the study were repeated. Summary statement of the patients studied in relation to age, gender and symptoms. Avoidance of a limited challenge group (to prevent excessively limiting the study group by excluding patients with clinical conditions that may cause false-negative or false-positive results). Adequate analysis of anatomical lesions (to allow evaluation of the test over the full anatomical spectrum of coronary artery disease). Adequate analysis for relevant chest pain syndromes (to ensure that the test was examined in patients with the common chest pain syndromes that usually lead to performance of the test)	**Verification bias**: avoidance of work-up bias	**Review bias**: avoidance of diagnostic review bias; avoidance of test review bias			**How were items chosen for inclusion on the scale?** Beginning with Ransohoff and Feinstein's discussion (1978),[66] the authors developed 7 methodological standards addressing important issues in diagnostic test research **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: angiography for coronary artery disease

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Radack, 1993[201] **Aim**: assess study quality **Type of scale**: quality score **Source of tool**: systematic review; authors' own tool	**Spectrum composition**: *Comparability of cases and controls*: required explicit description of the degree of similarity of members in the comparison groups for demographic factors (age and gender) **Population recruitment**: *Source of study subjects*: required adequate description of the study subjects and the study site, higher quality scores were assigned to subjects examined in primary care settings than to those evaluated at tertiary care centres *Selection of population*: required investigators to report the nature of how the study populations had been selected (e.g. consecutive case series, simple random sampling)	**Appropriate reference standard**: *Explicit definition of disease and marker*: defined polyps to be the disease and skin tags to be potential markers. Colonoscopies or post-mortem examination from a randomly selected population were considered to be acceptable diagnostic evaluations of the colon. Barium enemas, anoscopies and flexible sigmoidoscopies alone were not considered reference standard interventions **Verification bias**: *Full and comparable diagnostic evaluation for all subjects*: required full description of diagnostic intervention, appropriateness of its reproducibility and explicit statement regarding histological evaluation	**Review bias (test and diagnostic)**: *Blinded ascertainment of data*: required clear statement that endoscopist and skin examiners had been blinded to the results of the other's findings **Observer/instrument variability**: full description of diagnostic intervention, appropriateness of its reproducibility	**Appropriate results**: *Appropriate statistical methods*: had to report 2 × 2 tables to compute odds ratios, chi-squared and predictive values, explicit statement of statistical procedures to control for potential confounders and examination for dose–response effect indicating an assessment of the association between the number of skin tags and the presence of polyps		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: kappa = 0.88 for agreement between two appraisers **Topic area**: skin tags and colonic polyps

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Reid, 1995[26] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: methodological review; authors' own tool	**Spectrum composition**: spectrum should be specified. Standard met if at least 3 of the following 4 descriptors were provided: age distribution, gender distribution, summary of presenting clinical symptoms or disease stage, and inclusion criteria for study subjects	**Verification bias**: for cohort studies, standard met if all subjects were assigned to receive both diagnostic testing and reference standard verification either by direct procedure or by suitable clinical follow-up. In case–control studies, credit depended on whether the diagnostic test preceded or followed the reference standard procedure. If the diagnostic test preceded the reference standard, credit was given if disease verification was obtained for a consecutive series of study subjects regardless of their diagnostic test results. If the diagnostic test followed, credit was given if test results were stratified according to the clinical factor that evoked the reference standard procedure	**Review bias (test and diagnostic)**: for prospective cohort studies in which patients always received the diagnostic test first, credit was given if the reference standard procedures were evaluated independently. A statement about independence in interpreting both the test and the reference standard procedure was required for prospective studies in which the reference standard procedure was sometimes done before the diagnostic test and for case–control studies in which the test preceded the reference standard procedure. In case–control studies in which the diagnostic test followed disease verification, a statement was required to indicate an independent evaluation of the diagnostic test **Observer/instrument variation**: for tests requiring observer interpretation, at least some of the test subjects should have been evaluated for a summary measure of observer/instrument variation. For tests performed without observer interpretation, credit was given for a summary measure of instrument variability	**Precision of results**: this standard was met if SEs or confidence intervals, regardless of magnitude, were reported for test sensitivity and specificity or likelihood ratios. **Indeterminate test results**: to meet this standard a study had to report (1) all of the appropriate positive, negative and indeterminate results generated during evaluation of the diagnostic test, and (2) whether indeterminate results had been included or excluded when indices of accuracy were calculated **Analysis of subgroups**: this standard was fulfilled when results for indices of accuracy were cited for any pertinent demographic or clinical subgroup of the investigated population (e.g. symptomatic vs asymptomatic patients)		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: random sample of 24 studies reviewed blindly by 1 of 2 additional investigators (12 each): agreement 86% and 90% and kappa 0.72 and 0.75, respectively. Following revision of 1 criterion, subsequent blinded review of 12 studies by 2 observers showed perfect agreement **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Riegelman, 1996[214] **Aim**: guidelines for interpreting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: have the investigators included a broad enough cross-section of those with the disease to produce a realistic range of measurements for those with the disease?	**Appropriate reference standard**: have the investigators chosen the best available reference standard for defining which patients have the disease under study? **Test execution**: *Accuracy*: are the test results, on average, close to the true measure of the anatomical, physiological or biochemical phenomena? *Clinical or field accuracy*: under the usual conditions in which it is applied, has the test been shown to produce measurements that are close to the experimentally derived measurements? Has the purpose for use of the test been identified? **Normal defined**: has a reference interval been properly obtained to include a defined percentage, often 95% of those believed to be free of disease? Has outside the reference interval been distinguished from diseased? Has inside the reference interval been distinguished from disease free? Is the reference sample group used generally applicable, or are there identifiable reference sample groups with different reference intervals? Have those who applied the test recognised that the reference interval is a description of a presumably disease-free group and that changes within the reference interval for any one individual may be pathological? Has the reference interval been distinguished from desirable? Have the investigators justified moving the reference limits to accomplish specific diagnostic goals? If the test is designed to monitor progression of disease, has the change from previous levels been used to establish criteria for a positive test?	**Observer/ instrument variation**: *Precision*: do multiple repetitions of the test under the same conditions produce nearly identical results?	**Appropriate results**: how well does the test identify those with the disease? How high is its sensitivity? How often is it positive in disease? How well does the test identify those without the disease? How high is its specificity? How often is it negative in health? Have the sensitivity and specificity of the test been distinguished from the predictive value of a positive test and the predictive value of a negative test? If the test is designed to rule in a disease, has the test with the greater LR of a positive been identified as the better test to use to rule in disease? If so, has the relative importance of a false-negative result and a false-positive result been taken into account? If the test is designed to rule out a disease, has the test with the smallest LR of a negative been identified as the better test to use to rule out disease? If so, has the relative importance of a false-negative result and a false-positive result been taken into account? Have considerations of safety and cost been taken into account, as well as diagnostic performance when comparing tests? **Subgroup analyses**: has it been recognised that, despite the fact that in theory sensitivity and specificity are not affected by the probability of disease in the group being tested, they may be different for early versus more advanced disease?		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Rothwell, 2000[202] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: methodological review; authors' own tool	**Spectrum composition**: adequate detail of study population (age, gender, clinical presentation and indications for investigation) **Population recruitment**: prospective rather than retrospective study design. Patient selection based on a consecutive series or a random sample	**Test execution**: adequate detail of imaging techniques (sufficient for the study to be repeated) **Normal defined**: adequate detail of derivation of measurement of stenosis from images or data (sufficient for the study to be repeated)	**Review bias (test, diagnostic and clinical)**: blinded assessment of images **Clinical review bias**: blinding to clinical information **Observer/instrument variation**: adequate data on the reproducibility of measurements of stenosis (data on either intraobserver or interobserver agreement was accepted)	**Indeterminate test results**: inclusion of all investigations (i.e. patients with poor-quality imaging were not excluded)	**Sample size**: study powered according to a sample-size calculation	**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: imaging for carotid stenosis
Sackett, 1991[6] and Sackett, 1992[249] **Aim**: guidelines for interpreting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: has the diagnostic test been evaluated in a patient sample that included an appropriate spectrum of mild and severe, treated and untreated disease, plus individuals with different but commonly confused disorders? **Population recruitment**: was the setting for this evaluation, as well as the filter through which study patients passed, adequately described?	**Appropriate reference standard**: has there been an independent, blind comparison with a reference standard of diagnosis? **Test execution**: if the test is advocated as part of a cluster or sequence of tests, has its individual contribution to the overall validity of the cluster or sequence been determined? Have the tactics for carrying out the test been described in sufficient detail to permit their exact replication? **Normal defined**: has the term normal been defined sensibly as it applies to this test?	**Review bias**: has there been an independent, blind comparison with a reference standard of diagnosis? **Observer/instrument variation**: have the reproducibility of the test result (precision) and its interpretation (observer variation) been determined?	**Utility of test**: has the utility of the test been determined?		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported, assume not **Level of inter-rater reliability**: not reported **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Sackett, 2000[215] **Aim**: guidelines for interpreting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: was the diagnostic test evaluated in an appropriate spectrum of patients (like those in whom we would use it in practice)?	**Appropriate reference standard**: was there an independent blind comparison with a reference standard of diagnosis? **Verification bias**: was the reference standard applied regardless of the diagnostic test result?	**Review bias (test and diagnostic)**: was there an independent blind comparison with a reference standard of diagnosis? **Observer/instrument variability**: was the test (or cluster of tests) validated in a second, independent group of patients?			**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general
Sheps, 1984[27] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: methodological review; authors' own tool	**Disease prevalence/ severity**: is there any recognition of the influences of setting, prevalence and pre-test likelihoods on clinical utility?	**Appropriate reference standard**: is there a well-defined reference standard? The following were accepted as well defined: definitive histopathological diagnoses (autopsy, biopsy, surgery, etc.), standard diagnostic classification systems, or the results of other well-established diagnostic tests. The latter were considered well defined only if explicit criteria were given for when the target disease was said to be present **Normal defined**: are positive and negative clearly defined for the diagnostic test?	**Review bias (test and diagnostic)**: are the performance and interpretation of the diagnostic test explicitly stated to be blind? This required an explicit statement that those who performed and interpreted the diagnostic test were blind to the reference standard and vice versa	**Appropriate results**: are the terms sensitivity and specificity both used correctly, and are the calculations correct? Do the words predictive value or post-test likelihood or equivalent phrases appear in the article, and are the calculations correct? **Data table**: are the data clearly displayed in tabular form? Data had to be clearly presented as a 2×2 table		**How were items chosen for inclusion on the scale?** The authors state that the rationale for using these particular criteria comes both from the literature and from their desire to use the same criteria as those used in the 1982 study for the purposes of comparison **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported, but does not appear to have been **Level of inter-rater reliability**: on a random collection of 22 articles there was complete agreement for questions 6 and 7, and an observed agreement of 0.91 (kappa = 0.74) for question 5. Agreement was lower for the other questions, and was lowest for question 2, where there was an agreement of 0.77 (kappa = 0.55). After discussing the criteria the procedure was repeated with 26 articles and agreement improved, reaching 0.92 (kappa = 0.81) for question 2. A separate study[28] reports that on a random sample of 26 articles assessed independently by two reviewers the mean observed agreement for all criteria was 0.78, with a kappa score of 0.53. Criteria were reviewed and points of disagreement were discussed. A subsequent assessment of an additional 24 articles revealed a corresponding kappa score of 0.70 **Topic area**: general

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Sox, 1989[221] **Aim**: guidelines for reporting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: do the patients in the study population closely resemble the patients in the clinically relevant population? Is the study population described carefully enough to allow comparison to the clinically relevant population?	**Appropriate reference standard**: is the reference standard procedure an accurate measure of the true state of the patient? **Verification bias**: was an abnormal result on the index test a criterion for referring the patient for the reference standard test? **Normal defined**: choosing a definition of an abnormal result	**Review bias (test, diagnostic and clinical)**: if the index test or the reference standard test required visual interpretation, was the observer blinded to all other information about the patient? **Observer/instrument variation**: was interobserver disagreement measured?	**Analysis of subgroups**: were the true-positive rate and false-positive rate of the test measured in clinically relevant subgroups of patients?		**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general
Thornbury, 1991[220] **Aim**: guidelines for reporting study **Type of scale**: checklist **Source of tool**: original tool	**Spectrum composition**: *Referral bias*: the research team and referral environment should comprise physicians from a variety of medical practice environments in which potential patient subjects are first encountered by physicians	**Verification bias**: *Work-up bias*: all patients should receive both MR and the competing examination **Incorporation bias**: occurs when the reference standard diagnosis is affected by the imaging examination under study	**Review bias (test, diagnostic and clinical)**: readings of the competing examinations should be carried out blindly by multiple observers; test review bias occurs when the final diagnosis or the results of the comparison study are used in planning or interpreting the examination under study. **Clinical review bias**: observers should be blinded to clinical information at the time of imaging		**Sample size**: adequate numbers of patients must be obtained to provide statistical power to ensure that valid conclusions can be drawn	**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Assume not **Level of inter-rater reliability**: not reported **Topic area**: MRI

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
van der Wurff, 2000[203] **Aim**: assess study quality **Type of scale**: quality score **Source of tool**: systematic review; authors' own tool	**Spectrum composition**: A1. Description of study population, i.e. volunteers or patients, age, gender, etc. (8) **Inclusion criteria**: A2. Description of inclusion and exclusion criteria (7)	**Disease progression bias**: G. Test/retest procedure, description of time interval **Test execution**: D. Standardisation of test procedure: 1. Position of subject (3) 2. Position of examiner (2) 3. Description of palpation technique (position of hands of examiner) (3) 4. Description of neutralising simple exercises for low back and pelvis before or during the test procedure (2) 5. Information given to the subject about the test procedure (2) 6. Standardisation according to the original description of the test in the literature (referenced) (4) E. Selection of examiner 1. Description of the choice for experienced examiners (3) 2. Description of less experienced examiner (2) 3. Description of a consensus procedure (9) **Normal defined**: F. Standardised measurement of test outcome (5)	**Review bias**: H. Procedure of blinding: 1. Attempt to blind the examiner (2) 2. Subject not informed of outcome (1) 3. Results sealed, examiners could not see each other's findings (5)	**Appropriate results**: I. Descriptive statistics: frequencies and total agreement (10) J. Inferential statistics: Cohen's kappa or ICC **Dropouts**: B. Dropouts described, information from which group and with reason for withdrawal (5)	**Sample size**: C. Number of subjects: < 25 (0), >25 (3), >50 (6), >75 (10)	**How were items chosen for inclusion on the scale?** Not clearly reported. Authors state that they developed a criteria list according to the guidelines for meta-analysis evaluating diagnostic tests[130] and the method guidelines for systematic reviews by van Tulder.[139] Items that seemed to be irrelevant for reliability studies were dropped and more appropriate items were added **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Does not appear to have been **Level of inter-rater reliability**: the interexaminer reliability between the two reviewers was kappa = 0.63 **Topic area**: clinical tests of the sacroiliac joint

continued

Study details	Spectrum composition	Index test and reference standard		Data presentation	Research planning	Details of scale development
		Selection and execution	Interpretation			
Windeler, 1988[204] **Aim**: assess study quality **Type of scale**: checklist **Source of tool**: original		**Appropriate reference standard**: predefinition of a reference standard, and statements regarding its appropriateness **Test execution**: adequate description of test execution **Incorporation bias**: avoidance of use of pathological test results as part of the diagnosis of the disease (independence)	**Review bias**: interpretation of the test results without knowledge of the diagnoses and diagnosis without knowledge of the test results (blinding)	**Appropriate results**: sufficient data to produce a 2 × 2 table, presentation of sensitivity, specificity, predictive values, prevalence and miscellaneous terms. Number of true positives, false positives, true negatives and false negatives **Data table**: presentation of 2 × 2 table **Utility of test**: discussion of the relationship between the predictive values and the pre-test probability of disease	**Sample size**: power calculation before study started **Objectives**: development of a clear hypothesis before the start of the studies. Clear definition of the target illness	**How were items chosen for inclusion on the scale?** Not reported **Time taken to complete the scale**: not reported **Has the scale been rigorously developed?** Not reported **Level of inter-rater reliability**: not reported **Topic area**: general

Appendix 5

Members of the advisory panel

External advisors to the review

Mr David Moher (also member of the Delphi panel)
Director
Thomas C. Chalmers Centre for Systematic Reviews
Children's Hospital of Eastern Ontario Research Institute, Canada

Members of the advisory panel

Professor Colin Begg (also member of the Delphi panel)
Eugene W. Kettering Chair
Department of Epidemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center
New York, USA

Professor Patrick Bossuyt (member of Delphi panel only)
Head of Department
Department of Clinical Epidemiology and Biostatistics
Academic Medical Center
University of Amsterdam, The Netherlands

Jon Deeks (also member of the Delphi panel)
Senior Medical Statistician
Centre for Statistics in Medicine
University of Oxford, UK

Professor Constantine Gatsonis (also member of the Delphi panel)
Professor of Medical Science (Biostatistics) and Applied Mathematics
Director, Center for Statistical Sciences
Brown University, USA

Professor Les Irwig
Professor of Epidemiology
Department of Public Health and Community Medicine
University of Sydney, Australia

Dr Khalid Khan (also member of the Delphi panel)
Consultant
Birmingham Women's Hospital, UK

Dr Jeroen Lijmer (also member of the Delphi panel)
Universitair Medisch Centrum
Utrecht, The Netherlands

Professor Cynthia Mulrow (also member of the Delphi panel)
Deputy Editor
Annals of Internal Medicine

Dr Gerben Ter Riet (also member of the Delphi panel)
Clinical Epidemiologist
Department General Practice
Academic Medical Centre
University of Amsterdam, The Netherlands

Appendix 6

Delphi questionnaires

Delphi procedure – round 1

Part I: Instructions

The first round of the Delphi procedure consists of three sections:

- Instructions: questionnaire
- Delphi-1 questionnaire:
  - *general items*
  - *topic-specific items*
- Blank pages, which can be used to add general comments

Instructions: questionnaire

a. General items

The questionnaire contains 28 items concerning the quality of studies designed to evaluate diagnostic test performance. The items are classified into four categories, spectrum composition, index test and reference standard, analysis and research planning.

For each category you'll be asked to indicate whether or not this category should be included in the quality assessment tool. If you think a category should be included then you will be asked to indicate whether specific items within this category should be included in the criteria list. Please rate your impression on the five-point Likert scale.

To help with your decision-making process we have provided you with a summary of evidence from two systematic reviews which we have conducted. The first is a review of studies which provide evidence of the effects of bias on diagnostic test performance. For each source of bias we have summarised the number of studies providing each type of evidence, as follows:

Type of evidence	Number of articles identified
Empirical evidence of bias (E)	*n* =
Theoretical evidence of bias (T)	*n* =
Absence of bias (A)	*n* =

The second is a review of existing checklists used to assess the quality of diagnostic test evaluations. The proportion of studies covering each quality item was classified as follows:

Classification	Proportion of scales in which the item was included
I	75–100%
II	50–74%
III	25–49%
IV	0–24%

The number of studies providing each type of evidence of bias obtained from the first review (E, T and A) and the classification obtained from the second review (2) are provided for each item to help in the decision-making process. You can find the information in the following column in the questionnaire:

Evidence			
E	T	A	2
I	0	2	IV

In this example, one study found empirical evidence of bias, none provided theoretical evidence of bias and two studies found no evidence of bias. Less than 25% of the existing checklists for studies of diagnostic accuracy included this source of bias as an item.

Further details of the reviews, together with a more detailed description of each source of bias, are provided in the report which accompanies this document.

Topic-specific items

The reviews did not provide any evidence on the importance of topic-specific items. We have provided a list of possible topic areas. For each topic area you'll be asked to indicate whether or not this topic should be included in the quality assessment tool. Please rate your impression on the five-point Likert scale. If you think a topic should be included then you will be asked to suggest possible items for inclusion in the tool. You will also be asked to indicate whether you think that any of the general items would not be applicable to each topic area.

Target user group and purpose of the quality assessment tool

The criteria list will be used by reviewers conducting systematic reviews of diagnostic test evaluations to assess the quality of individual studies included in the review. For the purpose of this quality assessment tool, quality relates to both the internal and external validity of the study. Internal validity relates to the degree to which estimates of diagnostic accuracy produced in a study have not been biased as a result of study design, conduct, analysis or presentation. External validity refers to the degree to which the results of a study can be applied to patients in practice, and is affected by factors such as spectrum of disease, setting and other patient characteristics, how the diagnostic test was conducted and the reproducibility of the test.

It is anticipated that the tool will be used for conducting sensitivity analyses, to make recommendations for future research, as criteria for including studies in a review or in primary analyses and to be used in regression analyses. We do not anticipate that the tool be used to weight the meta-analysis. The tool will therefore not incorporate a quality score, but will be a list of items which will each be assessed as 'good', 'poor' or 'not reported'.

The final list should have the following properties:

- The list should enable the evaluation of the quality (internal and external validity) of the individual studies.
- The list should enable a qualitative indication of how likely a study is to produce biased estimates of test performance.
- The list should be practical it should consist of maximum 10 items, but with the option of additional topic-specific areas.

Verbal descriptions of the five-point Likert scale

- **Strongly disagree**: if, in your opinion, deviation on this item is unlikely to affect the results, or generalisability of the results, of a diagnostic test evaluation. This item definitely shouldn't be integrated in the criteria list.
- **Moderately disagree**: if, in your opinion, deviation on this item is not likely to affect the results, or generalisability of the results, of a diagnostic test evaluation.
- **Neutral**: if you're indifferent or if you don't know if deviation on this item has any association with the results, or generalisability of the results, of a diagnostic test evaluation.
- **Moderately agree**: if, in your opinion, deviation on this item might be related to the results of a diagnostic test evaluation, or the generalisability of the results, but the item is not essential to the list.
- **Strongly agree**: if, in your opinion, deviation on this item is equal to an inadequate diagnostic test evaluation, producing biased results, or affecting the generalisability of the results. This item should definitely be integrated in the criteria list.

In the decision whether or not the category and related items should be included in the criteria list, it might be helpful to ask yourself the following question: to what extent is the validity of conclusions of diagnostic test evaluations affected, if a study does not fulfil the category(s) and related item(s)?

Missing items and rephrasing

If you feel that we have omitted any items, you can *add items* in the 'Comments' section (on the pages on the left side of the domains and items). Please feel free to *add pro and contra arguments* in these sections. If you would like to suggest an item which you feel has been missed, please only suggest it if you would rate it as 'strongly agree' on the Likert scale. Please start your comment with the corresponding number of the category, section or related item you're referring to and whenever possible please add literature reference. If you sense that some items are expressed vaguely or ambiguously, you can *rephrase the item* in the comments section. If you need more space than provided in the 'Comments' section, please use the blank pages presented in part III of the questionnaire.

Part II: The Delphi-1 questionnaire

Category		Item	
C1	Spectrum composition	I1	Variation by clinical and demographic subgroups (spectrum composition)
		I2	Inclusion criteria
		I3	Distorted selection of participants
		I4	Disease prevalence/severity
C2	Index test and reference standard		
C2a	Selection and execution	I1	Absent or inappropriate reference standard
		I2	Change in technology of index test
		I3	Disease progression bias
		I4	Difference in test protocol (Test execution)
		I5	Difference in test protocol (Ref. execution)
		I6	Partial verification bias (Verification bias)
		I7	Differential verification bias (Verification bias)
		I8	Incorporation bias
		I9	Normal defined
		I10	Treatment paradox
C2b	Interpretation	I1	Review bias
		I2	Clinical review bias
		I3	Observer/instrument variation
C3	Analysis	I1	Appropriate results
		I2	Precision (sample size, variation by chance)
		I3	Inappropriate handling of uninterpretable/indeterminate/intermediate test results
		I4	Arbitrary choice of threshold value
		I5	Dropouts
		I6	Subgroups
		I7	Data table
		I8	Utility of test
C4	Research planning	I1	Sample size
		I2	Objectives
		I3	Protocol

C, category; I, item within a category.

Please mark the boxes that you want to select.

Category 1: Spectrum composition						strongly disagree	moderately disagree	neutral	moderately agree	strongly agree
C1	**Should the <u>category</u> 'Spectrum composition' be included in the criteria list?**							Go to C2a		Go to I
I.	**Should this <u>item</u> be included in the criteria list?**	Evidence								
		E	T	A	2					
1.	Was the spectrum of patients described in the paper and was it chosen adequately?	1 4	0	1	II					
2.	Were selection criteria described clearly?	na	na	na	IV					
3.	Was the method of population recruitment consecutive?	3	0	2	II					
4a.	Was the setting of the study relevant?	8	1	0	IV					
b.	Was disease prevalence and severity reported?									

Please add your **comments or rephrasing** of the items on the following page.

Category 2a: Index test and reference standard: selection and execution						strongly disagree	moderately disagree	neutral	moderately agree	strongly agree
C2a	**Should the <u>category</u> 'Index test and reference standard: Selection and execution' be included in the criteria list?**									
						Go to C2b			Go to I	
I.	**Should this <u>item</u> be included in the criteria list?**	Evidence								
		E	T	A	2					
1.	In light of current technology, was the reference test chosen appropriate to verify test results?	4	4	0	II					
2.	Is it possible that a change in the technology of the index test has occurred since this paper was published?	1	0	1	IV					
3.	Was there an abnormally long time period between the performance of the test under evaluation and the confirmation of the diagnosis with the reference standard?	0	0	1	IV					
4.	Was the execution of the index test described in sufficient detail to permit replication of the test?	1	0	1	III					
5.	Was the execution of the reference standard described in sufficient detail to permit replication of the test?				IV					
6.	Did the whole sample, or a random selection of the sample, receive verification using a reference standard of diagnosis?	1 7	3	3	II					
7.	Did all patients receive the same reference standard regardless of the index test result?	2	0	0						
8.	Were the results of the index test incorporated in the results of the reference standard?	0	0	0	IV					
9.	Was the cut-off value prespecified or acceptable in light of previous research?	na	na	na	III					
10.	Was treatment started based on the knowledge of the index test results before the reference standard was applied?	0	0	0	IV					

Please add your **comments or rephrasing** of the items on the following page.

Category 2b: Index test and reference standard: interpretation						strongly disagree	moderately disagree	neutral	moderately agree	strongly agree
C2b	**Should the <u>category</u> 'Index test and reference standard: Selection and execution' be included in the criteria list?**									
						Go to C3			Go to I	
I.	**Should this <u>item</u> be included in the criteria list?**	Evidence								
		E	T	A	2					
1a.	Were the index test results interpreted blind to the results of the reference standard?	4	0	1	I					
b.	Were the reference standard results interpreted blind to the results of the index test?									
2.	Were clinical data available when test results were interpreted?	7	0	1	IV					
3.	Are data presented on observer or instrument variation that could have affected the estimates of test performance?	8	0	0	III					

Please add your **comments or rephrasing** of the items on the following page.

Category 3: Analysis						strongly disagree	moderately disagree	neutral	moderately agree	strongly agree
C3	**Should the <u>category</u> 'Analysis' be included in the criteria list?**							Go to C4		Go to I
I.	**Should this <u>item</u> be included in the criteria list?**	Evidence								
		E	T	A	2					
1.	Were appropriate results presented (sensitivity, specificity, likelihood ratios, diagnostic odds ratios, predictive values) and were these calculated appropriately?	na	na	na	III					
2.	Was a measure of precision of the results presented (confidence intervals, standard errors)?	0	0	0	IV					
3.	Were uninterpretable/indeterminate/intermediate results reported and included in the results?	0	0	2	IV					
4.	Was the threshold value specified retrospectively based on analysis of the results?	0	0	0	IV					
5.	Were reasons for dropout from the study reported?	0	0	0	IV					
6.	Were subgroup analyses prespecified and clinically relevant?	na	na	na	IV					
7.	Were results presented in a 2 × 2 data table?	na	na	na	IV					
8.	Was any indication of the utility of the test provided?	na	na	na	IV					

Please add your **comments or rephrasing** of the items on the following page.

Category 4: Research planning				strongly disagree	moderately disagree	neutral	moderately agree	strongly agree
C4	**Should the category 'Research planning' be included in the criteria list?**			Process complete			Go to I	
I.	**Should this item be included in the criteria list?**	Evidence 1	Evidence 2					
1.	Was an appropriate sample size calculation performed and were sufficient patients included in the study?	na	III					
2.	Were study objectives clearly reported?	na	IV					
3.	Was there any evidence that a study protocol had been developed before the study was started?	na	IV					

Please add your **comments or rephrasing** of the items on the following page.

Part III: General comments

Delphi procedure – round 2

Part I: Instructions

The second round of the Delphi procedure consists of three sections:

- Decisions of the steering group
- Feedback from round 1
- Delphi-2 questionnaire.

Decisions of the steering group

Details are provided by the steering group (consisting of Penny Whiting, Jos Kleijnen, Anne Rutjes and Hans Reitsma) on how decisions were reached regarding which items to include in the final quality assessment tools. A number of other decisions were also made; these are detailed in this section.

Feedback

The feedback from the first round is to provide you with a summary of the responses of all panel members. This has been provided to you in two sections: a summary of the comments for each category and item, and a summary of the ratings for each category and item. Please read this carefully before moving onto the questionnaire and take this into consideration when rating each of the categories and items.

The second round questionnaire

The layout of the second round questionnaire is similar to that for the first round questionnaire; however, the way in which items are rated has been changed. Rather than rating each item on the five-point Likert scale, please indicate whether you think a category or item should be included or excluded from the quality assessment tool. Please consider the results from round 1, the comments from round 1, *and* the evidence provided for each item when deciding whether you think an item should be included in the final quality assessment tool.

Based on the results of the first round, items for which there were high levels of agreement were selected for inclusion in the final quality assessment tool. These items are shown at the top of the table for each category. Similarly, some items have been selected for exclusion from the list; these are shown at the bottom of the table for each category. These items should not be rated.

In addition to the section similar to that from the round 1 questionnaire, there is a second section to the questionnaire with a number of additional questions.

A scoring system for the final quality assessment tool has been proposed. You will be asked whether you agree with this system. If you do not agree with the proposed system, please suggest an alternative system together with an explanation of why you would prefer this.

For items that have been selected for inclusion in the final quality assessment and that have been rephrased you will be asked to indicate whether you agree with the rephrasing, or if not, to propose an alternative phrasing.

Instructions on how to use the quality assessment tool, together with descriptions of the included items, will be drawn up by the steering group. Copies of these together with the final version of the quality assessment tool will be sent to you as part of the third round of the Delphi procedure. You will be asked to state whether or not you approve of the instructions. For the current stage of the procedure you will be asked whether you object to this approach, and if so, to suggest an alternative approach.

Your support of the Delphi procedure will be sought. You will be asked whether you endorse the procedure so far, and if not, to make any suggestions for how it could be improved.

After the Delphi procedure has been completed the tool will be validated in a number of ways. Details of the proposed validation methods are presented. You will be asked whether you think these suggestions are appropriate, and if not to provide further suggestions of how you think the tool should be validated.

Decisions of the steering group

Missing values

Some categories were not scored by panel members. When categories had not been scored but items within the categories had been rated as 'moderately agree' or 'strongly agree' then the category was rated as 'strongly agree'. In one case, one item within a category was missing; this was rated as 'neutral'.

Rephrasing of items

All comments regarding rephrasing were considered and items have been rephrased taking these into account. Additionally, items have been rephrased so that if the answer to each question is 'yes' then this indicates that the study is unlikely to be biased in respect of this item.

Selection of items for inclusion in final quality assessment tool

All categories/items rated as 'strongly agree' by at least six/eight of the Delphi panel members were selected for inclusion in the tool. These items will not be rated as part of this round, and will be included in the final quality assessment tool. This differs slightly from the rules laid out as part of round 1 where it was stated that all items rated 'strongly agree' or 'moderately agree' by all panel members would be included in the final list. It was felt that it was more important to focus on items for which a high proportion of the panel members voted 'strongly agree'. This change decreased the number of items to be rerated.

Selection of items for exclusion from the final quality assessment tool

Categories/items which were not rated as 'strongly agree' by at least one panel member were excluded. These items will not be rated as part of this round and are excluded from the final quality assessment tool. This is in agreement with the rules laid out as part of the round 1 questionnaire for how items would be selected for removal from the list. It was also stated that items rated as 'neutral' or lower by all panel members would be removed from the list. There were no items for which this occurred.

All other items

All other items showed disagreement among panel members. Therefore, these items will be given a second chance and will be rated for inclusion in the final quality assessment tool.

Definition of adequate/appropriate/abnormally, etc.

The definitions of what is meant by these terms will vary according to the specific topic area in which the quality assessment tool is used. It is therefore not possible to be more specific about these items. Users of the tool will have to define these for their specific topic areas before using the tool. This will be explained in the instructions accompanying the tool.

Feedback: summary of responses to round 1

Of the 11 people invited to take part in the Delphi procedure, eight took part in the first round procedure and returned completed questionnaires.

Comments from panel members

Overall remarks

- I think the last line of the table on page 5 should read "Absence of evidence of bias", not "Absence of bias".
- No item should address two or more subitems at once, since such constructions can make consistent scoring impossible, e.g. "Was disease prevalence and severity reported?" What to score if prevalence is reported but severity in an insufficient fashion?
- I would prefer to separate threats to the validity of the study – which depends necessarily on the study question – from lack of usability due to poor reporting, from poor or high quality planning.
- I have expressed reservations about a one-size-fits-all quality list for studies on diagnostic accuracy, and I still do, after reading the (excellent) documents and the provisional checklist. The Delphi procedure is not going to solve this. This procedure may result in consensus, but it will not solve the conceptual issues. On second thought, quality of reporting *is* important for a systematic review. It may limit the way in which [sentence stopped here].

Comments on the categories and the corresponding items

Category 1: Spectrum composition

General comments

- Spectrum composition depends on the study question, which should be described in the paper. Consecutive sampling is not a necessity, provided the sampling mechanism is described and appropriate for the study question.
- Quality of *reporting* differs from the quality of the study design itself. Poor reporting makes the study less useful, without further information from the authors. It does not make the study invalid.
- I would avoid using the word "spectrum". In clinical evaluations, a term like "patient cohort" would be more specific and understandable.
- Enrolling consecutive patients is often one of the ways to avoid selection bias. Why not phrase the question in a way that allows other ways by which such bias can be avoided. Example: in a screening study, it may not be practical to all people who come in to be screened. Thus the researchers may be forced to find a way of choosing among a large number of possible participants while they still avoid selection bias.
- Spectrum composition has to do with clinical heterogeneity. It is *not* directly relevant to bias (which produces methodological heterogeneity). But in practice it probably does not make too much difference as many reviewers cannot make the distinction. On balance, I might be persuaded to include this item.

Specific comments
1. Was the spectrum of patients described in the paper and was it chosen adequately?

- Replace "adequately" with "appropriately"
- "Was the spectrum of patients described in the paper and was it chosen adequately?" Adequacy of choosing can only be assessed if description is available. I suggest to use "Was the spectrum of patients chosen adequately?"
- "was it chosen adequately" is not a good phrase. Entirely non-objective. I think that the issue should be whether test performance statistics are calculated separately for groups likely to have different values – i.e. it's an issue of first whether the spectrum is described and secondly whether subgroups likely to have different values have been combined as one or separated. Also separate spectrum of diseased and non-diseased for case–control type studies.

3. Was the method of population recruitment consecutive?

- consecutive or random

4a. Was the setting of the study relevant?

- Relevant to what?
- Were the referral stages through which patients reached this study described?
- This should be part of 1.

Category 2a: Index test and reference standard: selection and execution
Specific comments
1. In light of current technology, was the reference standard chosen appropriate to verify test results?

- In the light *of the study question* – not necessarily appropriate technology. The appropriateness of the reference standard depends on the purpose of the study. Follow-up for example may be appropriate, compared to pathology or invasive testing.
- Is the reference standard likely to produce results close to the true disease status?

2. Is it possible that a change in the technology of the index test has occurred since this paper was published?

- This may be answerable by YES too often. Perhaps the issue is that the technology investigated is the technology used in the setting that wants to use the review's results. Variation in technology may be interesting in a review context.

3. Was there an abnormally long time period between the performance of the test under evaluation and the confirmation of the diagnosis with the reference standard?

- "Abnormally" will cause problems.
- Is the time period between reference and index tests short enough to be reasonably sure that disease states did not change between the two tests?

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

- Replication of the "article"?

  Clarification by the steering group: This item is not meant to evaluate whether or not the original article could be replicated. It intends to enable the investigation of possible heterogeneity due to differences in test protocol in the different studies included in the review.

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

- The term "reference standard" or "reference information" seems more broadly applicable than "reference test".
- Why use the term "gold standard" here – stick to "reference standard" (see also Item 8: ref test).
- This has more to do with the *design* of the study, an aspect that is absent from this list.
- Gold standard should be reference standard.

  Clarification by the steering group: A new item has been added to category 1 (spectrum composition). This item asks "What was the study design?" The possible answers for this are:
  1. Diagnostic cohort, index test performed first.
  2. Diagnostic cohort, reference standard performed first.
  3. Diagnostic case–control study.

  This item will have implications for which other items on the quality assessment tool are relevant, and also for how certain items are answered. For example, in a diagnostic cohort study in which the reference standard is performed first or in a diagnostic case–control study, verification bias will not apply.

Instructions explaining this will be produced to accompany the quality assessment tool.

Item 1 in category 1 has been rephrased to read "has an appropriate spectrum of patients been included?" Whether this is answered 'yes' will be different for each study design: for a diagnostic case–control study the answer will be 'yes' if the control group contains participants with other diseases similar to the target condition and is not just a 'healthy control' group, and if the cases are not just those with the most severe form of the disease but with an appropriate spectrum of disease severity. For a diagnostic cohort study it will be answered as 'yes' if the cohort composition in terms of disease prevalence and severity, gender, age, etc., is similar to the situation in which the test will be used in practice. Further details of this will be explained in the instructions accompanying the final quality assessment tool.

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

- Should be split: (1) all verified or not (2) verified using the same reference standard.
- It is not always possible to use the same reference standard for all patients, but there should be some room to allow for use of more than one reference standard as long as they are both independent of the index test.
- Incorporates 6 to some extent.

8. Were the results of the index test incorporated in the results of the reference standard?

- Suggested rephrasing: "Was the reference standard evaluated with knowledge of the result of the index test?" Also, how is this different from the issues in category 2b?
- Contained in other items.
- Please note that sometimes (rarely, I admit) the reference information may actually include the results of the test. Example: In study evaluating the accuracy of core needle biopsy, if the result of the biopsy (the 'test' being evaluated) shows cancer, then the reference standard would also be cancer.

9. Was the cut-off value prespecified or acceptable in light of previous research?

- This issue belongs in "analysis".
- Two aspects: prespecified/acceptable; depends also on the study question.
- Should be in the next section.
- Were cut-off values used in interpreting results derived independently of the results of the current study?

10. Was treatment started based on the knowledge of the index test results before the reference standard was applied?

- The question about treatment (#10) does not seem to belong to this section, at least not by itself. If the intention is to discern whether treatment affected the reference standard information, more than one question may be needed.
- You may consider rephrasing this statement with use of the term "treatment paradox", which exemplifies that use of an effective treatment in light of the index test may make the test look bad (inaccurate) in the study as the results of reference standard will be modified due to treatment (particularly if the index is accurate and the treatment is effective).
- Or vice versa.

Category 2b: Index test and reference standard: interpretation

Specific comments

1. Were the index test results interpreted blind to the results of the reference standard?

- Replace "blind to the" with "without knowledge of the".

2. Were clinical data available when test results were interpreted?

- Needs to be separated for the two tests.
- 1, 2 & 7 could be simplified to a question of whether sufficient data are provided to include a study in a systematic review (requires crude data values). But this is an issue of study reporting not internal/external validity, so I wouldn't mind if it were dropped.
- As long as the same data is available to both group of interpreters, bias will not be likely.

3. Were data presented on observer or instrument variation that could have affected the estimates of test performance?

- Item 3 is ambiguous. One item is whether or not data were presented on reproducibility. A second is whether or not reproducibility is/was high enough to put a limit on measures of accuracy.
- Was the reliability of the tests (especially index

test) demonstrated to be similar to that achieved in clinical practice?

Category 3: analysis
Specific comments
1. Were appropriate results presented (sensitivity, specificity, likelihood ratios, diagnostic odds ratios, predictive values) and were these calculated appropriately?

- This is very vague. What does "appropriate" mean?
- Item 1 is ambiguous – please split.

3. Were uninterpretable/indeterminate/ intermediate results reported and included in the results?

- Remove "and included in the results"
- Should not say both "reported and included" on the results – because what is meant by "included" is unclear
- Item 3 is also ambiguous. Uninterpretable results (if any) should be reported, so the reader may make his or her own calculations. The way in which authors do so [sentence stopped here].

4. Was the threshold value specified retrospectively based on analysis of the results?

- Hard to know how you would figure this out.
- Item 4 duplicates the threshold item mentioned previously.
- Is duplication of C2a(9).

5. Were reasons for dropout from the study reported?

- Numbers and reasons.

6. Were subgroup analyses prespecified and clinically relevant?

- Hard to know how you would figure this out.
- Item 6 is a double one. Provided a study has sufficient power, a subgroup analysis may make a study more useful, but this also depends on the purpose of the systematic review.
- Relates to questions on spectrum and could be incorporated as suggested.

7. Were results presented in a 2 × 2 data table?

- One would want the raw numbers to be able to construct 2 × 2 tables, but is it reasonable to expect an actual *table* in the article?
- Main issue is that no matter the presentation used, reviewers are put into the position to draw the required tables: 2 × 2, 3 × 2, 4 × 3, etc., or the distribution of test results in stem and leaf diagrams (formats that contain the raw data). Perhaps different requirements should be considered for continuous and ordinal test results.
- The best way of reporting the results when one makes a SR is presenting raw numbers, a 2 × 2 table – or *n* by *n* table – anyway.
- My main concern here is that this is phrased as if all test evaluations involve binary test results. Obviously this is not the case and hence the criteria need to be expanded to include ROC studies and beyond.

8. Was any indication of the utility of the test provided?

- Define "utility". The accompanying article is vague on this point also.
- I did not understand Item 8. What utility?

  Clarification by the steering group: The utility of the test refers to how useful the test will be in practice.

Category 4: Research planning
General comments
- Generally I think if a study was based on a research protocol then this is a big plus, but my impression of this literature is that this would be really quite rare (except perhaps for studies of screening tests).
- Very generally, for a reviewer any issues concerned with the planning of the published studies are irrelevant. The reviewer has to work with what materialised.
- All items relate to the quality of planning – important, but in itself not related to external or internal validity.

Specific comments
1. Was an appropriate sample size calculation performed and were sufficient patients included in the study?

- Sample size may be a surrogate, but not a real factor implicating bias. Moreover, there is no agreed approach to sample size estimation in test accuracy studies.
- Was an appropriate sample size calculation performed and were sufficient patients included in the study? The use of "and" should be avoided: two items in one question.
- Sample size depends on the study question. Absence of sample size does not invalidate a study.

- The notion of adequate sample size after the study was done is tricky.
- Don't know how to do sample size calculation.

2. Were study objectives clearly reported?

- Piffle! – These studies are really poorly written, but doesn't mean they aren't useful or reasonable quality – there are more direct ways to assess this than by asking questions like this.

3. Was there any evidence that a study protocol had been developed before the study was started?

- Would be nice – but again unlikely to be a good question to discriminate between good and bad studies (can't remember when I last read a study like this).

Feedback: Summary of panel members' rating of each category and item

Category/item		Likert score[a]				
		1	2	3	4	5
Included categories and items (do not rate)						
C1	*Spectrum composition*[b]	*0*	*0*	*1*	*1*	*6*
C2a	*Index test and reference standard: selection and execution*	*0*	*0*	*0*	*1*	*7*
I1	In light of current technology, was the reference test chosen appropriate to verify test results?	0	0	0	1	7
I6	Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis?	0	0	2	0	6
C2b	*Index test and reference standard: interpretation*	*0*	*0*	*0*	*0*	*8*
I1a	Were the index test results interpreted blind to the results of the reference standard?	0	0	0	0	8
I1b	Were the reference standard results interpreted blind to the results of the index test?	0	0	0	1	7
I2	Were clinical data available when test results were interpreted?	0	1	0	1	6
C3	*Analysis*	*0*	*0*	*0*	*1*	*7*
I3	Were uninterpretable/indeterminate/intermediate results reported and included in the results?	0	1	0	1	6
Items to be rerated						
C1	*Spectrum composition*[b]					
I1	Was the spectrum of patients described in the paper and was it chosen adequately?	0	1	0	1	5
I2	Were selection criteria described clearly?	0	0	3	1	3
I3	Was the method of population recruitment consecutive?	0	0	4	2	1
I4a	Was the setting of the study relevant?	0	1	2	1	3
I4b	Was disease prevalence and severity reported?	0	1	0	2	4
C2a	*Index test and reference standard: selection and execution*					
I2	Is it possible that a change in the technology of the index test has occurred since this paper was published?	0	2	1	4	1
I3	Was there an abnormally long time period between the performance of the test under evaluation and the confirmation of the diagnosis with the reference standard?	0	1	0	5	2
I4	Was the execution of the index test described in sufficient detail to permit replication of the test?	0	0	2	1	5
I5	Was the execution of the reference standard described in sufficient detail to permit replication of the test?	0	0	2	1	5
I7	Did all patients receive the same reference standard regardless of the index test result?	0	0	1	2	5
I8	Were the results of the index test incorporated in the results of the reference standard?	1	0	1	3	3
I9	Were the cut-off values prespecified or acceptable in light of previous research?	0	1	3	1	3
I10	Was treatment started based on the knowledge of the index test results before the reference standard was applied?	0	1	4	0	3
C2b	*Index test and reference standard: interpretation*					
I3	Were data presented on observer or instrument variation that could have affected the estimates of test performance?	0	1	2	1	4

continued

Category/item		Likert score[a]				
		1	2	3	4	5
C3	*Analysis*					
I1	Were appropriate results presented and were these calculated appropriately?	1	0	3	0	4
I2	Was a measure of precision of the results presented?	1	0	2	2	3
I4	Was the threshold value specified retrospectively based on analysis of the results?	0	0	4	2	2
I5	Were reasons for dropout from the study reported?	0	0	1	3	4
I6	Were subgroup analyses prespecified and clinically relevant?	0	0	4	2	2
I7	Were results presented in a 2 × 2 table?	0	2	3	0	3
C4	*Research planning*[c]	*0*	*1*	*4*	*1*	*2*
I1	Was an appropriate sample size calculation performed?	0	1	0	1	1
I2	Were study objectives clearly reported?	0	0	0	0	3
I3	Was there any evidence that a study protocol had been developed before the study started?	0	1	0	0	2
Excluded items (do not rate)						
C3	*Analysis*					
I8	Was any indication of the utility of the test provided?	0	0	7	1	0

[a] Likert score: 1, strongly disagree; 2, moderately disagree; 3, neutral: 4, moderately agree; 5, strongly agree.
[b] The category 'Spectrum bias' was only rated by seven panel members, as one member rated this category as neutral.
[c] The category 'Research planning' was only rated by three panel members, as five panel members rated this category as neutral or less.
C, category; I, item within a category.

PLEASE READ THE FEEDBACK FROM ROUND 1 (COMMENTS AND RATINGS) BEFORE COMPLETING THE DELPHI-2 QUESTIONNAIRE. PLEASE ALSO TAKE INTO ACCOUNT THE EVIDENCE PROVIDED FROM THE SYSTEMATIC REVIEWS.

We would like to remind you that the intended use of final quality assessment tool is for reviewers conducting systematic reviews of diagnostic test evaluations to assess the quality of individual studies included in the review.

As a reminder, the evidence from the systematic reviews is summarised as follows:

Type of evidence	Number of articles identified
Empirical evidence of bias (E)	*n* =
Theoretical evidence of bias (T)	*n* =
Absence of bias (A)	*n* =

and for 2:

Classification	Proportion of scales in which the item was included
I	75–100%
II	50–74%
III	25–49%
IV	0–24%

Part II: The Delphi-2 questionnaire

Included items are shown at the top of the tables; excluded items at the bottom – these items should not be rated as part of this round. For all other items please mark whether you think this item should be included or excluded.

A scoring system has been devised for each item. This is presented at the end of each category. You will be asked to state whether you agree with the scoring system, and if not to suggest an alternative system and to explain why you would prefer this.

At the end of the section relating to selection of items for the quality assessment tool you will find a number of additional questions. Please answer these as directed.

Category 1: Spectrum composition

Items to be rerated:							
I.	**Should this item be included in the criteria list?**	**Evidence**				**Exclude**	**Include**
		E	**T**	**A**	**2**		
New item	*What was the study design?*						
1.	Was the spectrum of patients selected appropriately?	14	0	1	II		
2.	Were selection criteria clearly described?	na	na	na	IV		
3.	Was a random or consecutive sample of patients included in the study?	3	0	2	II		
4a.	Was the setting of the study relevant?	8	1	0	IV		
b.	Was disease prevalence reported?						
c.	Was disease severity reported?						

The following scoring system is proposed:

Items 1–4: Yes/no/not stated

	Yes	No
Do you agree with this scoring system?		

New item: diagnostic cohort, index test performed first
diagnostic cohort, reference standard performed first
diagnostic case control

	Yes	No
Do you agree with this scoring system?		

If no, please suggest an alternative scoring system on the following page and explain why you would prefer this system.

Please add your **comments or rephrasing** of the items on the following page.

Category 2a: Index test and reference standard: selection and execution

Included items
1. Is the reference standard likely to produce results close to the true disease state?
6. Did the whole sample, or a random selection of the sample, receive verification using a reference standard of diagnosis?

Items to be rerated:

I.	Should this item be included in the criteria list?	Evidence				Exclude	Include
		E	T	A	2		
2.	Is it possible that a change in the technology of the index test has occurred since this paper was published?	1	0	1	IV		
3.	Is the time period between reference standard and index test short enough to be reasonably sure that disease status did not change between the two tests?	0	0	1	IV		
4.	Was the execution of the index test described in sufficient detail to permit replication of the test?	1	0	1	III		
5.	Was the execution of the reference standard described in sufficient detail to permit replication of the test?				IV		
7.	Did patients receive the same reference standard regardless of the index test result?	2	0	0	II		
8.	Did the results of the index test form part in the reference standard?	0	0	0	IV		
9.	Was the definition of a 'normal' test result reported?	na	na	na	III		
10.	Was treatment started based on the knowledge of the index test results before the reference standard was applied (treatment paradox)?	0	0	0	IV		

The following scoring system is proposed:

Items 1, 2, 3, 8, 10: Yes/no/not stated

	Yes	No
Do you agree with this scoring system?		

Items 6, 7: Yes/no/not stated/not applicable (depending on study design)

	Yes	No
Do you agree with this scoring system?		

Items 4, 5, 9: Yes/no

	Yes	No
Do you agree with this scoring system?		

If no, please suggest an alternative scoring system on the following page and explain why you would prefer this system.

Rephrasing

Item 1:
Previous phrasing: In light of current technology, was the reference standard chosen appropriate to verify test results?
Proposed phrasing: Is the reference standard likely to produce results close to the true disease state?

	Yes	No
Do you accept the proposed phrasing?		

Item 6:
Previous phrasing: Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?
Proposed phrasing: Did the whole sample, or a random selection of the sample, receive verification using a reference standard of diagnosis?

	Yes	No
Do you accept the proposed phrasing?		

Please add your **comments or rephrasing** of the items on the following page:

Category 2b: Index test and reference standard: interpretation

Included items
1a. Were the index test results interpreted without knowledge of the results of the reference standard?
b. Were the reference standard results interpreted without knowledge of the results of the index test?
2. Were clinical data available when index test results were interpreted?

Items to be rerated:

I.	Should this item be included in the criteria list?	Evidence				Exclude	Include
		E	T	A	2		
3a.	Were data presented on observer variation?	8	0	0	III		
b.	Were data presented on instrument variation?						

The following scoring system is proposed:

Items 1a, 1b, 2: Yes/no/not stated

	Yes	No
Do you agree with this scoring system?		

Items 3a, 3b: Yes/no

	Yes	No
Do you agree with this scoring system?		

If no, please suggest an alternative scoring system on the following page and explain why you would prefer this system.

Rephrasing

Item 1a:
Previous phrasing: Were the index test results interpreted blind to the results of the reference standard?
Proposed phrasing: Were the index test results interpreted without knowledge of the results of the reference standard?

	Yes	No
Do you accept the proposed phrasing?		

Item 1b:
Previous phrasing: Were the reference standard results interpreted blind to the results of the index test?
Proposed phrasing: Were the standard results interpreted without knowledge of the results of the index test?

	Yes	No
Do you accept the proposed phrasing?		

Item 2:
Previous phrasing: Were clinical data available when test results were interpreted?
Proposed phrasing: Were clinical data available when index test results were interpreted?

	Yes	No
Do you accept the proposed phrasing?		

Please add your **comments or rephrasing** of the items on the following page.

Category 3: Analysis

Included items	
3.	Were uninterpretable/indeterminate/intermediate results reported?

Items to be rerated:							
I.	**Should this item be included in the criteria list?**	**Evidence**				**Exclude**	**Include**
		E	T	A	2		
7.	Were sufficient results presented to calculate an $n \times n$ (e.g. 2×2) data table?	na	na	na	IV		
1.a	Were appropriate results presented?	na	na	na	III		
b	Were results calculated appropriately?						
New item	Were sufficient data provided to include the study in a systematic review?						
2.	Was a measure of precision of the results presented (confidence intervals, standard errors)?	0	0	0	IV		
4.	Were threshold values used in interpreting results derived independently to the results of the current study?	0	0	0	IV		
5.	Were both numbers and reasons for drop-out from the study reported?	0	0	0	IV		
6a.	Were subgroup analyses prespecified?	na	na	na	IV		
b.	Were subgroup analyses clinically relevant?						

Excluded items	
8.	Was any indication of the utility of the test provided?

The following scoring system is proposed:

Items 1b, 4, 6a: Yes/no/not stated

	Yes	No
Do you agree with this scoring system?		

Items 1a, 2, 3, 5, 6b, 7, new item: Yes/no

	Yes	No
Do you agree with this scoring system?		

Rephrasing:

Item 3:
Previous phrasing: Were uninterpretable/indeterminate/intermediate results reported and included in the results?
Proposed phrasing: Were uninterpretable/indeterminate/intermediate results reported?

	Yes	No
Do you accept the proposed phrasing?		

If no, please suggest an alternative scoring system on the following page and explain why you would prefer this system.

Please add your **comments or rephrasing** of the items below:

Category 4: Research planning

Items to be rerated:							
						Exclude	**Include**
C4	**Should the category 'Research planning' be included in the criteria list?**						
I.	**Should this item be included in the criteria list?**	**Evidence**				**Exclude**	**Include**
		E	**T**	**A**	**2**		
1.	Were sufficient participants included in the study?	na			III		
2.	Were study objectives clearly reported?	na			IV		
3.	Was there any evidence that a study protocol had been developed before the study was started?	na			IV		

The following scoring system is proposed:

Items 1, 2, 3: Yes/no

	Yes	No
Do you agree with this scoring system?		

If no, please suggest an alternative scoring system on the following page and explain why you would prefer this system.

Please add your **comments or rephrasing** of the items on the following page.

Additional questions

	Yes	No
1. Would you like to see a number of 'key items' highlighted in the quality assessment tool?		

If yes, please list those items which you would like to see highlighted:

Category number	Item number	Brief description of item

	Yes	No
2. Do you endorse the Delphi procedure so far?		

If no, please give details of the aspects of the procedure which you do not support and list any suggestions you have for how the procedure could be improved:

	Yes	No
3. As part of the third round, instructions on how to complete the quality assessment will be provided to you. As we do not want to ask you to invest too much time, the instructions will be drawn up by the steering group. In the third round you will only be asked if you support the instructions and if not, what you would like to change. Do you agree with this procedure?		

If no, please suggest an alternative approach:

4. Validation of the tool

The following methods have been planned to validate the quality assessment tool once the Delphi procedure is complete. Please indicate which steps you think are appropriate (mark as yes). If you do not think a step should be included, please provide a short description of why you think it should be excluded. Finally, please provide any further suggestions you have on how the tool should be validated.

Validation step	Yes	No
1. The revised instrument will be piloted by three raters on a sample of published studies in order to identify any problems in clarity or application of the items. The items will then be reworded and instructions clarified if necessary.		
2. The following step has been used in similar procedures (Jadad *et al*., 1996): The frequency of endorsement is "the proportion of people who give each response alternative to an item". Items where one alternative has a very high or low endorsement may be eliminated as they do not help to discriminate between good and poor studies. We **do not** plan to include this step. Do you agree with this?		
3. The consistency or reliability can be measured by the degree to which different individuals agree on the scientific quality of a set of papers. Three groups of raters, researchers, clinicians and 'others', will be randomly allocated to open or blinded assessment of the same set of studies. The raters will be asked to independently assess the quality of the report, using the developed instrument, with no additional training in how to score the items. Intraclass correlation coefficients (ICCs) and their 95% confidence intervals will be used to measure the agreement between raters. Items with ICCs greater than 0.50 will be considered to be sufficiently reliable, and those scoring greater than 0.65 represent a high level of agreement.		
4. The instrument will be adjusted based on the outcome of the above steps		
5. A regression analysis will be used to investigate associations between study characteristics and estimates of diagnostic accuracy in primary studies, as combined in existing systematic reviews. The methods used to conduct this analysis will be similar to the approach taken by Lijmer *et al*. (1999).		
6. The tool will be piloted in a number of diagnostic reviews. Current projects planned include reviews of tests for TB, UTI in children, appendicitis, prediction of pre-eclampsia, and prediction of preterm labour.		

References

Jadad AR, Moore A, Carroll D, *et al*. Assessing the quality of reports of randomised clinical trials: is blinding necessary? *Control Clinical Trials* 1996;**17**:1–12.

Lijmer JG, Mol BW, Heisterkamp S, *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6.

Part III: General comments

Delphi procedure – Round 3

Part I: Instructions

The third round of the Delphi procedure consists of four sections:

- Decisions of the steering group
- Feedback from round 2
- Delphi-3 questionnaire
  - phrasing of included items
  - items to be rerated
  - additional questions
- Background document to accompany the quality assessment tool.

Decisions of the steering group

Details are provided by the steering group (consisting of Penny Whiting, Jos Kleijnen, Anne Rutjes and Hans Reitsma) on how decisions were reached regarding which items to include in the final quality assessment tool. A number of other decisions were also made; these are detailed in this section.

Feedback

The feedback from the second round is to provide you with a summary of the responses of all panel members. This has been provided to you in two sections: a summary of the comments for each category and item, and a summary of the ratings for each category and item.

Delphi-3 questionnaire

Phrasing of included items

A number of the items selected for inclusion in the quality assessment tool have been rephrased based on the responses received from round 2. You will be asked to indicate whether you agree with the final wording of each of the items, both those which have been rephrased and those which have not. You will be given the opportunity to suggest alternatives if there is any phrasing with which you strongly disagree.

Items to be rerated

There was still disagreement regarding a number of items. You will be asked to rerate these items, indicating whether you think each item should be included in the quality assessment tool.

Based on some of the comments received from the feedback from round 2, four additional generic questions have been proposed. You will be asked to indicate whether you think each of these four questions should be included in the quality assessment tool.

Additional questions

You will be asked to indicate whether you agree with the scoring proposed for the quality assessment tool.

Based on some of the feedback from round 2 it was clear that we had not provided adequate descriptions of some of the proposed validation methods. These methods have been explained in more detail, and you will again be asked to indicate whether you agree with the proposed validation methods.

Your support of the Delphi procedure, given the goal of the project, will again be sought. You will be asked whether you endorse the procedure so far, and if not, to make any suggestions for how it could be improved.

We would like to know whether you used the evidence provided from the systematic reviews, and from feedback from the previous rounds, in making your decisions of which items to include in the quality assessment tool.

Based on some of the comments received from the feedback from round 2, we have added a number of questions regarding the possible development of topic- and design-specific items. You will be asked to indicate whether you would like to see the development of topic- and/or design-specific items. If you would like to see such items included then you will be asked to indicate whether you would like to see these developed through a second Delphi procedure, and if so whether you would like to be a panel member for this procedure.

Background document to accompany the diagnostic quality assessment tool

A background document which provides details of exactly what is meant by each of the questions included in the diagnostic quality assessment tool has been produced. This document also explains how each item should be scored, and gives details of situations in which it may be appropriate to exclude certain items from the quality assessment tool. We would like to have your comments on this document.

Decisions of the steering group

Rephrasing of items

All comments regarding rephrasing were considered and items have been rephrased taking these into account.

Selection of items for inclusion in final quality assessment tool

All categories/items rated as 'include' by at least

seven/eight of the Delphi panel members were selected for inclusion in the tool. Items scored 'include' by six/eight of the panel members will be rerated as part of round 3. All other items were removed from the tool.

Feedback: summary of responses to round 2

Of the ten people invited to take part in the Delphi procedure, eight took part in the second round procedure and returned completed questionnaires.

Comments from panel members

Category 1: spectrum composition

New item: What was the study design?

- The new item is very general …
- Consider a study in which a set of patients with a particular form of cancer has been selected on the basis of a clinical workup. The patients are then imaged with two competing modalities before undergoing surgery to determine the true disease. The order of the two imaging studies could be determined systematically or at random. How would such studies be classified in the schema proposed above?
- Consider another study in which people are coming to be screened for a particular disease. Two screening tests are available, each participant undergoes both tests, and the order of the tests is randomised. Where would this scenario fit?

Item 1: Was the spectrum of patients selected appropriately?

- Too vague.
- Was the spectrum of patients selected appropriately *with respect to the aim of the review*.
- The "aim of the review" seems crucial unless this is part of the instructions accompanying the tool. I wonder what the theoretical variation of review questions is. If all review questions have the format: "What is the diagnostic accuracy of test T in patients suspected of (suspicion should be operationalised) Y, (compared to the accuracy of test U)?", appropriate selection automatically pertains to the reviewer's perspective.
- Was spectrum "appropriate" – were the participants described in adequate detail (disease, severity, presenting symptoms, referral process)?
- This item requires a subjective interpretation of the text, which is often difficult as knowledge of the clinical problem is necessary.

Item 2: Were selection criteria described clearly?

- Change "clearly described" to "defined".

Item 4a: Was the setting of the study relevant?

- Was … relevant *in view of the review's objective(s)*.
- This item requires a subjective interpretation of the text, which is often difficult as knowledge of the clinical problem is necessary.

Item 4c: Was disease severity reported?

- Was *the distribution* of disease severity reported (in sufficient detail)?

General comments

- I'm having big problems with the options which are listed here. There are very important issues in this section which we risk losing without care.

Category 2a: Index test and reference standard: selection and execution

Item 1: Is the reference standard likely to produce results close to the true disease state?

- Do not agree with proposed phrasing. There is not always a "true disease state"; this is a very subjective statement, prone to observer bias.
- Do not agree with the proposed phrasing. The true disease state is not necessarily the target condition.

Item 2: Is it possible that a change in the technology of the index test has occurred since this paper was published?

- Not always applicable (think of lab tests …).

Item 5: Was the execution of the reference standard described in sufficient detail to permit replication of the test?

- Was the execution … to permit *its replication*. Not always the reference is *a* test, e.g. clinical follow-up.

Item 8: Did the results of the index test form part of the reference standard?

- I draw attention to my previous comment about the existence of situations in which it is natural to include the test results in the definition of the reference information. These are not very common situations, but do come up. In such cases the interpretation of the answer to item 8 above will be misleading.

Category 2b: Index test and reference standard: interpretation

Item 1a: Were the index test results interpreted without knowledge of the results of the reference standard?
Item 1b: Were the reference standard results interpreted without knowledge of the results of the index test?

- 1a, 1b. I wonder whether these two can be combined into one item by using the phrase vice versa or the like.

Item 2: Were clinical data available when the results of the index test were interpreted?
- Were clinical data available when either index or reference test results were interpreted?
- Again, the use of the framework of "index test" and "reference test" is problematic. The general setting includes one or more diagnostic test under evaluation and a reference standard.
- Isn't the issue *which* clinical data were available? If the index test is meant to replace other (clinical) tests, they should not be available. In other situations the test results that normally precede the index test, if any, should be available.
- Yes, but what about the reference test? Does knowledge of clinical data make a difference, or are we assuming that the reference test definition states that clinical data are or are not used? I can see that they are different, but this is not discussed in the notes provided.

Rephrasing of item 2
- Don't agree.
- Don't know what is meant by clinical data?
- Is the intent to assess whether the person who interpreted index test results had knowledge of other factors, either clinical, historical, laboratory that could have biased the interpretation?
- The intent of this question is unclear.

Category 3: Analysis
General comments
- What you believe should go here depends on what you think a reviewer should do with the data. If 2 × 2 data are available then it is irrelevant as to whether sensitivity and specificity are appropriately calculated – the review will redo this. If 2 × 2 data are not available then the article will not be included thus what might be thought of as a quality issue becomes an issue of inclusion. RCT scales do not usually ask whether data is presented – I don't think we should waste questions doing this.
- Items 1, 7 and new item are part of assessing "eligibility" for an SR, not assessing study quality.

Item 1b: Were results calculated appropriately?
- 1b can only be answered if 7 scores a YES, but then become superfluous since the reviewer can do the calculations him/herself.

New item
- The new item is of course logically necessary, but I argue that it is not relevant here.

Item 2 – Was a measure of precision of the results presented?
- 2 should not be a problem if 7 scores a YES. If 7 scores NO, item 2 seems pretty worthless. Any useful reporting of item 2 should enable the reviewer to make the article fulfil item 7.

Item 4: Were threshold values used in interpreting results derived independently to the results of the current study?
- "… independently to" or "of " or "from"?
- Item 4 addresses the problem of validation in independent data sets. What to do with studies that used split sample techniques and the like to 'validate' their thresholds? Is not the historically first study forced to use data-dependent thresholds? Cannot the estimates of the accuracy parameters based on data-dependent thresholds be considered as maximum values? Suppose study 1 (1999) defines its threshold conditional on the data it produced. Suppose study 2 (2000) uses study 1's threshold and performs a second analysis using its own data to suggest a better threshold or produces an ROC curve? What are criteria for arriving at a correct threshold?

Item 5: Were both numbers and reasons for dropouts reported?
- Should be split into two.
- This is about two things, early comment says one item per question.

Item 6a and b
- These are important for critical appraisal of individual studies but can we assume subgroup analyses are part of the SR protocol, thus not relevant?

Category 4: Research planning
Item 2: Were study objectives clearly reported?
- Were study objectives specified?

General comments
- Can we add two generic questions:
- Experience of looking at test evaluation studies is that they can generate 'bad' ways of doing studies in previous unthought of ways – many of which might be critical.
- One about study methods?
  Are there other aspects of the design and conduct of this study which cause concern about whether or not it will correctly estimate test accuracy?
- One about topic/test specific issues?
  Are there special issues concerning patient selection and the conduct of tests which might invalidate test results?

Scoring system

General

- Don't agree for all items: use not stated or unclear
- I prefer 'Don't know' over not stated because 'not stated' does not always follow logically from the item-questions, whereas 'don't know' always applies and includes 'uncertain' and 'partly' which may occur quite often when YES or NO are too strong an expression for what has been reported.
- Don't agree: use unclear rather than not stated.
- Not sure that the distinction of 'yes/no/not stated' and 'yes/no/not stated/not applicable' and 'yes/no' really matters.

Study design

- Don't agree with scoring for study design.
- Don't agree with this classification (for study design) unless it is better defined.
- I am not sure how to interpret the terminology proposed by the steering group. The language of "index test" and "reference test" does not fit the reality of many studies.
- Far too general open question, with limited number of options.
- The scoring system for the study design should contain more than three options, as there are at least two more alternatives. One could rephrase the question to "Was the study design described clearly?"

Endorse Delphi procedure

- No – the process has been very clear and straightforward. However, I fundamentally believe that it is not possible to develop a reliable discriminatory diagnostic assessment tool that will apply to all, or even the majority of diagnostic test studies.
- I have explained my reservations on previous occasions:
  - There is a distinction between:
  - Quality of reporting (do you find the information you need)
  - The usability of the paper (fit between study purpose and the reader's questions)
  - The potential for bias (fit between study design and study purpose)
  - Lack of applicability (in general)
  - There will never be a one size-fits all quality list, I am afraid.
- ? – the process seems rather driven by the original items – it is not easy to suggest new items/combining items, etc. I think that some expert discussion may have been of value at an early stage.

  Comment from steering group: The objective of this project was not to produce a tool to cover everything, but to produce a quality assessment tool that can be used to assess the quality of primary studies included in systematic reviews. We appreciate that different aspects of quality will be applicable to different topic areas and for different study designs. We see this section of the quality assessment tool as the generic part of what in practice may be a more extensive tool incorporating design- and topic-specific items. Therefore we propose to develop the quality assessment tool further by developing a section on topic- and design-specific items: for some topics/designs items included in the final quality assessment tool will not be applicable and certain items not included in the tool may be important. Once the generic section of the tool has been finalised, we propose to run a second Delphi procedure to develop topic- and design-specific items. We have added questions regarding your opinions on this suggestion at the end of the Delphi-3 questionnaire.

Validation

Step 2

- I do not understand it.

Step 3

- About validation step 3: what is meant by open and blinded assessment? The selection mechanism of the three groups of users seems crucial. In addition, numbers should be large enough to calculate narrow CIs around the ICCs within each group to allow for the advice that the tool is reliable only in the hands of e.g. clinicians. Compliance with the instructions is important. I am not sure whether future users should be advised to learn how to use the tool on a sample of publications before they start to use it in their formal review.
- I understand kappa coefficients but not ICC. Difficult to endorse your arbitrary values.

Step 4

- It is also not clear how the instrument will be "adjusted" as specified in step 4.

Step 5

- This analysis only looks at impact of quality components on DOR, and is limited accordingly. Why not look at evidence of differences in sensitivity and specificity as well?

  Comment from the steering group: From the comments received it is clear that we have not

provided sufficient information on some of the proposed validation methods. We have therefore added a section to the end of the round Delphi-3 questionnaire giving more details of the methods which were rated as '?' by at least one reviewer, and have again provided you with the opportunity to give your opinions on these methods.

Feedback: Summary of panel members' rating of each category and item

Category/Item		Decision		
		In	Ex	?
Included items				
C1	*Spectrum composition*			
1	Was the spectrum of patients selected appropriately?			
C2a	*Index test and reference standard: selection and execution*			
11	Is the reference standard likely to correctly classify the target condition?			
13	Is the time period between reference standard and index test short enough to be reasonably sure that disease status did not change between the two tests?	7	1	0
16	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?			
17	Did patients receive the same reference standard regardless of the index test result?	8	0	0
18	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	8	0	0
C2b	*Index test and reference standard: interpretation*			
11a	Were the index test results interpreted without knowledge of the results of the reference standard?			
11b	Were the reference standard results interpreted without knowledge of the results of the index test?			
12	Were clinical data available when test results were interpreted?			
C3	*Analysis*			
13	Were uninterpretable/indeterminate/intermediate results reported?			
15	Were both numbers and reasons for dropout from the study reported?	8	0	0
Items to be rerated				
C1	*Spectrum composition*			
12	Were selection criteria clearly described?	6	1	1
14c	Was disease severity reported?	6	1	1
C2a	*Index test and reference standard: selection and execution*			
14	Was the execution of the index test described in sufficient detail to permit replication of the test?	6	2	0
15	Was the execution of the reference test described in sufficient detail to permit replication of the test?	6	2	0
C3	*Analysis*			
14	Were threshold values used in interpreting results derived independently to the results of the current study?	6	2	0
Excluded items				
C1	*Spectrum composition*			
New	What was the study design?	5	3	0
13	Was a random or consecutive sample of patients included in the study?	5	2	1
14a	Was the setting of the study relevant?	0	8	0
14b	Was disease prevalence reported?	5	3	0

continued

Category/Item		Decision		
		In	Ex	?
C2a	*Index test and reference standard: selection and execution*			
I2	Is it possible that a change in the technology of the index test has occurred since this paper was published?	2	6	0
I9	Was the definition of a 'normal' test result reported?	5	3	0
I10	Was treatment started based on the knowledge of the index test results before the reference standard was applied (treatment paradox)?	5	3	0
C2b	*Index test and reference standard: interpretation*			
I3a	Were data presented on observer variation?	3	5	0
I3b	Were data presented on instrument variation?	3	5	0
C3	*Analysis*			
I7	Were sufficient results presented to calculate an $n \times n$ (e.g. 2×2) data table?	4	4	0
I1a	Were appropriate results presented?	3	5	0
I1b	Were results calculated appropriately?	3	5	0
New	Were sufficient data provided to include the study in a systematic review?	2	7	0
I2	Was a measure of precision of the results presented?	4	4	0
I6a	Were subgroup analyses prespecified?	1	7	0
I6b	Were subgroup analyses clinically relevant?	2	6	0
I8	Was any indication of the utility of the test provided?			
C4	*Research planning*[a]	4	4	0
I1	Were sufficient participants included in the study?	0	4	0
I2	Were study objectives clearly reported?	4	0	0
I3	Was there any evidence that a study protocol had been developed before the study started?	3	1	0

[a] The category 'Research planning' was only rated by four panel members, as two panel members rated this category as 'exclude'. Shaded areas indicate items which were 'included' or 'excluded' from the quality assessment tool based on the results of round 1 and which were therefore not rerated as part of round 2.
C, category; I, item within a category.

Feedback: Summary of panel members' response to each additional question

Question	Response		
	Yes	No	Unclear
1. Would you like to see a number of 'key items' highlighted in the quality assessment tool?	0	8	0
2. Do you endorse the Delphi procedure so far?	5	1	2
3. Procedure for instructions on completing tool	8	0	0
Validation step			
1. Piloting by three raters	8	0	0
2. Exclusion of frequency of endorsement step. Do you agree with this?	7	0	1
3. Assessment of consistency and reliability of the instrument	6	0	2
4. The instrument will be adjusted based on the outcome of the above steps	7	0	1
5. Regression analysis	5	1	2
6. The tool will be piloted in a number of diagnostic reviews	8	0	0

Part II: The Delphi-3 questionnaire

Phrasing of included items

Please indicate whether you agree with the phrasing for items selected for inclusion in the quality assessment tool, and if not please explain why and suggest an alternative phrasing using the space provided on the following page. Items shown in italics have been rephrased based on feedback from round 2.

Item		**Do you agree with the phrasing?**	
		Yes	**No**
1.	*Was the spectrum of patients representative of the patients who will receive the test in practice?*		
2.	*Is the reference standard likely to classify the target condition correctly?*		
3.	Is the time period between reference standard and index test short enough to be reasonably sure that disease status did not change between the two tests?		
4.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?		
5.	Did patients receive the same reference standard regardless of the index test result?		
6.	*Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?*		
7a.	Were the index test results interpreted without knowledge of the results of the reference standard?		
7b.	Were the reference standard results interpreted without knowledge of the results of the index test?		
8.	*Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?*		
9.	Were uninterpretable/indeterminate/intermediate results reported?		
10.	*Were withdrawals from the study accounted for?*		

Items to be reassessed for inclusion

Please indicate whether you think each of the following items should be included in the quality assessment tool. When making your decision please consider the feedback from round 2, the evidence from the systematic reviews, and the items already selected for inclusion in the quality assessment tool. Please only mark items as 'include' if you feel very strongly that these items should be included in the quality assessment tool.

Items to be rerated		**Evidence**				**Include**	**Exclude**
I	**Should this item be included in the criteria list?**	**E**	**T**	**A**	**2**		
1.	Were selection criteria clearly defined?	na	na	na	IV		
2.	Was disease severity reported?	8	1	0	IV		
3.	Was the execution of the index test described in sufficient detail to permit replication of the test?	1	0	1	III		
4.	Was the execution of the reference standard described in sufficient detail to permit its replication?				IV		
5.	Were threshold values used in interpreting results derived independently to the results of the current study?	0	0	0	IV		

Based on the feedback from round 2 additional generic questions have been proposed. Please indicate whether you would like to see these items included in the quality assessment tool:

Additional items		**Include**	**Exclude**
I	**Should this additional item be included in the criteria list?**		
1.	Are there other aspects of the design of this study which cause concern about whether or not it will correctly estimate test accuracy?		
2.	Are there other aspects of the conduct of this study which cause concern about whether or not it will correctly estimate test accuracy?		
3.	Are there special issues concerning patient selection which might invalidate test results?		
4.	Are there special issues concerning the conduct of tests which might invalidate test results?		

Additional questions

1. All items included in the quality assessment tool will be scored as 'yes', 'no' or 'unclear'.

	Yes	No
Do you agree with the scoring proposed for the quality assessment tool?		

If no, please suggest an alternative below:

2. Validation methods
There was some confusion regarding exactly what was meant by some of the validation methods outlined in the Delphi-2 questionnaire. Further details are provided on steps which were rated as unclear by at least one person. For each of these steps please indicate whether you agree with this step and if not please explain why. Everyone that responded to the Delphi-2 questionnaire agreed with validation steps 1 and 6. These should not be reassessed as part of this round.

Validation step	Yes	No
Accepted procedures		
1. The revised instrument will be piloted by three raters on a sample of published studies in order to identify any problems in clarity or application of the items. The items will then be reworded and instructions clarified if necessary.		
6. The tool will be piloted in a number of diagnostic reviews. Current projects planned include reviews of tests for TB, UTI in children, appendicitis, prediction of pre-eclampsia, and prediction of preterm labour.		
Procedures requiring clarification which should be reassessed		
2. The following step has been used in similar procedures (Jaded *et al.*, 1996): The frequency of endorsement is "the proportion of people who give each response alternative to an item". It is calculated by dividing the number of times each item is scored by the maximum possible number of times each of the items could have been scored, multiplied by 100. Items where one alternative has a very high or low endorsement may be eliminated as they do not help to discriminate between good and poor studies. Items which score similarly on good and poor quality reported would not help to discriminate between these. We **do not** plan to include this step. Do you agree with this?		

continued

Validation step	Yes	No
3. The consistency or reliability can be measured by the degree to which different individuals agree on the scientific quality of a set of papers. Three groups of raters: researchers, clinicians and 'others', will assess the same set of studies. We plan to include around five people in each group of raters. The raters will be asked to independently assess the quality of the report, using the quality assessment tool and the background document, with no additional training on how to score the items. Intraclass correlation coefficients (ICCs) (for a more detailed description of these please see Shrout and Fleiss) and their 95% confidence intervals will be used to measure the agreement between raters. Although any choice of cut-off will be arbitrary we plan to use the same cut-offs as used by Jadad *et al*. Items with ICCs greater than 0.50 will be considered to be sufficiently reliable, and those scoring greater than 0.65 represent a high level of agreement.		
4. The instrument will be adjusted based on the outcome of the above steps. Although the actual items included in the quality assessment tool will not be changed, the phrasing and instructions accompanying each item may need to be adjusted if it appears that these are not being interpreted and applied as intended.		
5. A regression analysis will be used to investigate associations between study characteristics and estimates of diagnostic accuracy in primary studies, as combined in existing systematic reviews. The methods used to conduct this analysis will be similar to the approach taken by Lijmer *et al*. A regression model adapted from the summary receiver operating characteristic curve developed for meta-analyses of diagnostic tests will be fitted to the data. The logarithm of the diagnostic odds ratio (DOR) computed for a single study will be modelled as the dependent variable. Dependent variables for the intercept and slope of the curve will be fitted for each meta-analysis. Covariates for each methodological feature of the new assessment scale will be added simultaneously to this model. The resulting parameter estimates of the covariates can be interpreted after antilogarithm transformation as relative diagnostic odds ratios. They indicate the diagnostic performance of a test in studies failing to satisfy the methodological criterion, relative to its performance in studies with the corresponding feature. If the relative diagnostic odds ratio is larger than 1, studies not satisfying the criterion yield larger estimates of the diagnostic odds ratio than studies with this feature. This process will be carried out for several meta-analyses with relatively large numbers of included studies, for both diagnostic accuracy outcomes and therapeutic and/or patient outcomes. In addition to the analysis done by Lijmer *et al*., looking at the associations between characteristics and diagnostic odds ratios, the association of these characteristics with sensitivity and specificity will also be investigated.		

References

Jadad AR, Moore A, Carroll D, *et al*. Assessing the quality of reports of randomised clinical trials: is blinding necessary? *Control Clin Trials* 1996;**17**:1–12.

Lijmer JG, Mol BW, Heisterkamp S, *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6.

Shront PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**2**:420–8.

Please add any comments on the following page.

	Yes	No
3. Given the comments from the steering group do you endorse the Delphi procedure so far?		

If no, please give details of the aspects of the procedure which you do not support and list any suggestions you have for how the procedure could be improved:

	Yes	No
4. Did you use the evidence provided from the systematic reviews to help make decisions on which items to include in the quality assessment tool?		

If you did not use this information, please explain why you chose not to use it:

	Yes	No
5. Did you use the evidence provided from the feedback from round 1 to help make decisions on which items to include in the quality assessment tool?		

If you did not use this information, please explain why you chose not to use it:

	Yes	No
6. Would you like to see the development of topic-specific items in addition to the generic quality assessment tool?		

Please use the space below to add any comments that you have:

	Yes	No
7. Would you like to see the development of design-specific items in addition to the generic quality assessment tool?		

Please use the space below to add any comments that you have:

	Yes	No
8. If you would like to see the development of topic- and/or design-specific items would you like to see this done via a Delphi procedure?		

Please use the space below to add any comments that you have:

	Yes	No
9. If you would like to see the development of topic- and/or design-specific items via a Delphi procedure, would you like to be part of the Delphi panel?		

Please use the space below to add any comments that you have:

10. Background document

The Background document to accompany the quality assessment tool is provided below. We have decided to call the tool the QUADAS (Quality Assessment of Diagnostic Accuracy Studies) tool. The aim of this document is to present the tool, to explain what is meant by each of the items included in the tool, to explain situations in which these items may not apply, and to give guidance on how to score the items.

Please read this carefully and add any comments that you have as you read through it, or if you would like to make further comments provided please use the additional blank page at the end of the document.

Delphi procedure – round 4

Part I: Instructions

The fourth round of the Delphi procedure consists of five sections:

- Decisions of the steering group
- Feedback from round 3
- Background document to accompany QUADAS
- Presentation of final QUADAS tool
- Comments.

We anticipate that this will be the final round of the Delphi procedure. This round does not contain a separate questionnaire although you will be given the opportunity to comment on any aspects of the tool. If there are any comments which have a significant impact on the tool (e.g. rephrasing of items) or on the background document then revisions will be made and you will be sent the revised tool and background document.

Decisions of the steering group

Details are provided by the steering group (consisting of Penny Whiting, Jos Kleijnen, Anne Rutjes and Hans Reitsma) on how decisions were reached regarding which items to include in the final quality assessment tool. A number of other decisions were also made; these are detailed in this section.

Feedback

The feedback from the third round is to provide you with a summary of the responses of all panel members. This has been provided to you in two sections: a summary of the comments for each question, and a summary of the ratings for each question.

Presentation of the final QUADAS tool

Based on the feedback to the last round the final version of the QUADAS tool has been developed. You will be given the opportunity to comment on the items included in the tool and also the phrasing of the items.

Background document to accompany QUADAS

The background document has been revised based on comments received from round 3 of the Delphi procedure. The additional items selected for inclusion in QUADAS have been added to the background document. You will be given the opportunity to comment on the revised background document.

Comments

You will be given a final chance to comment on any features of the QUADAS tool or background document which you are not happy with. For example, if you feel strongly that an important item has been omitted from the tool, or if you are still unhappy with the phrasing of any of the items.

Decisions of the steering group

Rephrasing of items

All comments regarding rephrasing were considered and items have been rephrased taking these into account.

Selection of items for inclusion in final quality assessment tool

All categories/items rated as 'include' by at least seven/nine of the Delphi panel members were selected for inclusion in the tool. All other items were removed from the tool.

Development of topic- and design-specific elements

More than half of panel members indicated that they would like to see the development of design- and topic-specific criteria, and of these four stated that they would like to see this done via a Delphi procedure. The development of these elements will therefore take place after the generic section of the tool has been validated.

Feedback: Summary of responses to round 3

Of the ten people invited to take part in the Delphi procedure, nine took part in the third round procedure and returned completed questionnaires.

Comments from panel members

Suggestions for rephrasing of items:

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

- Item 1 OK to have different spectrums, as practices are different – just need to have the chosen spectrum well described so one knows how to generalise.

Item 3: Is the time period between reference standard and index test short enough to be reasonably sure that disease status did not change between the two tests?

- Replace "disease status" with target condition.

Item 6: Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?

- "Was the reference standard interpreted without knowledge of the index test (i.e. the index test did not form part of the reference standard)?"

- On reflection exclude item 6 as it is contained in 7a.
- There is no need to repeat comments I have already made on my disagreement with the wording "index test", etc., examples of situations when the results of a test may form part of the reference information.

  Comment from the steering group: Based on the comments received there seems to be some confusion regarding the difference between items 6 and 7. Item 7 relates to blinding – whether the investigators were aware of the results of the index test when interpreting the results of the reference standard and vice versa. Item 6 relates to whether the index test actually formed a part of the reference standard. For example, a study investigating MRI for the diagnosis of MS could have a reference standard composed of clinical follow-up, CSF analysis and MRI. In this case the index test forms part of the reference standard and item 6 would be scored as 'no'. If the same study used a reference standard of clinical follow-up and the results of the MRI were known when the clinical diagnosis was made but were not specifically included as part of the reference standard then item 6 would be scored as 'yes' but item 7b would be scored as 'no'.

Item 8: Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?
- Disagree with phrasing.
- It is more useful to know which clinical data were available. This allows for replication of the study and makes it more easy to decide whether or not the use of clinical data might have biased results.

Item 9: Were uninterpretable/indeterminate/intermediate results reported?
- Suggest: "Were uninterpretable/intermediate results reported?"
- Suggest: "Were uninterpretable/indeterminate/intermediate test results reported?"

Item 10: Were withdrawals from the study accounted for?
- Suggest: "Were withdrawals from the study explained?"

General comments

Comments on phrasing of items
- These items may be difficult to understand for non-specialists. Explanation is required.

Comments on items to be rerated

Item 5: Were threshold values used in interpreting results derived independently to the results of the current study?
- This is a minor issue, as not for many test evaluations the threshold values were predefined by the factory or by for example the radiologic society.

Comments on suggested additional items
- I agree with the general principle that a study can still be 'bad' or 'useless' for reasons not mentioned in the checklist, but what is the use of including such generic questions in an item-by-item checklist? This makes little sense to me, because it brings back unsystematic 'expert judgment'.
- It seems to me that the additional questions are too broadly worded and overlap with many of the specific items listed already in the questionnaire. Perhaps what you are looking for is a "fatal flaw" in the design, conduct, or interpretation of the results of the study. Hopefully such fatal flaws will not be too common and could be accommodated with a single question.
- These can be rephrased to one general additional item; are there other aspects of this study which cause concern … accuracy?

1. Scoring of items
- How about including "not applicable" – this holds for many items (think of all the questions related to "interpretation", 7a and 7b and lab tests, for example).

2. Validation method

Step 2
- ? Why not ? Please calculate!

Step 5
- Odds ratio (DOR) computed – not optimal.
- The regression approach only looks for systematic bias. Some of these features may introduce bias in different directions in different reviews, which on average may appear to "cancel each other out" – care and thought is needed.
- Step 5 is interesting, but does not fall under "validation" as such, does it? There is no description of the consequences of this analysis for the checklist itself.

3. Given the comments from the steering group, do you endorse the Delphi procedure so far?
- Addition of these latest stages has been of value.

- Please be as clear as possible with respect to the purpose of the instrument. All my reservations still apply.

4. Did you use the evidence provided from the systematic reviews to help make decisions on which items to include in the quality assessment tool?
- No new stuff inside.
- Too busy!
- Depends, much of it has muddled quality of reporting with study quality, and thinking and evidence has progressed, so some of it is out of date. Also it's quite clear that much 'copying' of items has occurred over time, so frequency of inclusion is not necessarily a good measure of value.
- Yes, but only slightly.

5. Did you use the evidence provided from the feedback from round 1 to help make decisions on which items to include in the quality assessment tool?
- I carefully read all texts, but I am not sure to what extent I really used that information.
- To some extent – very little actually. I found the 'Delphi' character less useful for non-quantitative questions (as all of these are).

6. Would you like to see the development of topic-specific items in addition to the generic quality assessment tool?
- Yes, but by the people doing each review. Maybe some examples can be given, but reviewers should be encouraged to apply their own brains to these issues.
- What do you mean by "develop topic- and design-specific items"?

7. Would you like to see the development of design-specific items in addition to the generic quality assessment tool?
- Unclear of what they would be – RCTs are obvious, but do you mean only for diagnostic accuracy?

QUADAS background document

There were a number of changes which panel members had marked in the text; these have been incorporated into the background document. One additional comment is highlighted below.

- It may be possible to construct a table explaining the coding instruction for 'yes', 'no' and 'unclear' as in CRD report 4 appendix 2.

Summary of responses

1. Phrasing of included items

Item		Do you agree with the phrasing?	
		Yes	**No**
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?	8	1
2.	Is the reference standard likely to classify the target condition correctly?	9	0
3.	Is the time period between reference standard and index test short enough to be reasonably sure that disease status did not change between the two tests?	8	1
4.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	9	0
5.	Did patients receive the same reference standard regardless of the index test result?	9	0
6.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	7	2
7a.	Were the index test results interpreted without knowledge of the results of the reference standard?	9	0
7b.	Were the reference standard results interpreted without knowledge of the results of the index test?	9	0
8.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	7	2
9.	Were uninterpretable/indeterminate/intermediate results reported?	7	2
10.	Were withdrawals from the study accounted for?	8	1

Items to be reassessed for inclusion

Items to be rerated		Include	Exclude
I	**Should this item be included in the criteria list?**		
1.	Were selection criteria clearly defined?	7	2
2.	Was disease severity reported?	4	5
3.	Was the execution of the index test described in sufficient detail to permit replication of the test?	7	2
4.	Was the execution of the reference standard described in sufficient detail to permit its replication?	7	2
5.	Were threshold values used in interpreting results derived independently to the results of the current study?	3	6

Additional items		Include	Exclude
I	**Should this additional item be included in the criteria list?**		
1.	Are there other aspects of the design of this study which cause concern about whether or not it will correctly estimate test accuracy?	3	6
2.	Are there other aspects of the conduct of this study which cause concern about whether or not it will correctly estimate test accuracy?	4	5
3.	Are there special issues concerning patient selection which might invalidate test results?	3	6
4.	Are there special issues concerning the conduct of tests which might invalidate test results?	3	6

Additional questions

Question	Response		
	Yes	**No**	**Unclear**
1. Do you agree with the scoring proposed for the quality assessment tool?	9	0	0
2. Validation step			
Exclusion of frequency of endorsement step. Do you agree with this?	7	2	0
Assessment of consistency and reliability of the instrument	9	0	0
The instrument will be adjusted based on the outcome of the above steps	9	0	0
Regression analysis	8	1	0
3. Given the comments from the steering group do you endorse the Delphi procedure so far?	8	0	1
4. Did you use the evidence provided from the systematic reviews to help make decisions on which items to include in the quality assessment tool?	7	2	0
5. Did you use the evidence provided from the feedback from round 1 to help make decisions on which items to include in the quality assessment tool?	6	3	0
6. Would you like to see the development of topic-specific items in addition to the generic quality assessment tool?	5	4	0
7. Would you like to see the development of design-specific items in addition to the generic quality assessment tool?	5	4	0
8. If you would like to see the development of topic- and/or design-specific items would you like to see this done via a Delphi procedure?	4	1	0
9. If you would like to see the development of topic- and/or design-specific items via a Delphi procedure, would you like to be part of the Delphi panel?	4	0	0

Part II: Presentation of the QUADAS tool

Items shown in italics have been rephrased based on the feedback from round 3. Items shown in bold are items which have been added to the tool based on the results of round 3.

	Item
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 classify the target condition correctly?
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)?
8a.	**Was the execution of the index test described in sufficient detail to permit replication of the test?**
8b.	**Was the execution of the reference standard described in sufficient detail to permit its replication?**
9a.	Were the index test results interpreted without knowledge of the results of the reference standard?
9b.	Were the reference standard results interpreted without knowledge of the results of the index test?
10.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?
11.	*Were uninterpretable/intermediate test results reported?*
12.	*Were withdrawals from the study explained?*

Part III: General comments

Please use this space to add any final comments that you have on any aspect of the QUADAS, including the items included in the tool and the phrasing of included items. If you have any comments on the QUADAS background document you can either add these to the document itself or list them below.

Health Technology Assessment Programme

Prioritisation Strategy Group

Members

Chair,
Professor Tom Walley,
Director, NHS HTA Programme, Department of Pharmacology & Therapeutics, University of Liverpool

Professor Bruce Campbell, Consultant Vascular & General Surgeon, Royal Devon & Exeter Hospital

Professor Shah Ebrahim, Professor in Epidemiology of Ageing, University of Bristol

Dr John Reynolds, Clinical Director, Acute General Medicine SDU, Radcliffe Hospital, Oxford

Dr Ron Zimmern, Director, Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge

HTA Commissioning Board

Members

Programme Director,
Professor Tom Walley,
Director, NHS HTA Programme, Department of Pharmacology & Therapeutics, University of Liverpool

Chair,
Professor Shah Ebrahim,
Professor in Epidemiology of Ageing, Department of Social Medicine, University of Bristol

Deputy Chair,
Professor Jenny Hewison,
Professor of Health Care Psychology, Academic Unit of Psychiatry and Behavioural Sciences, University of Leeds School of Medicine

Dr Jeffrey Aronson
Reader in Clinical Pharmacology, Department of Clinical Pharmacology, Radcliffe Infirmary, Oxford

Professor Ann Bowling, Professor of Health Services Research, Primary Care and Population Studies, University College London

Professor Andrew Bradbury, Professor of Vascular Surgery, Department of Vascular Surgery, Birmingham Heartlands Hospital

Professor John Brazier, Director of Health Economics, Sheffield Health Economics Group, School of Health & Related Research, University of Sheffield

Dr Andrew Briggs, Public Health Career Scientist, Health Economics Research Centre, University of Oxford

Professor Nicky Cullum, Director of Centre for Evidence Based Nursing, Department of Health Sciences, University of York

Dr Andrew Farmer, Senior Lecturer in General Practice, Department of Primary Health Care, University of Oxford

Professor Fiona J Gilbert, Professor of Radiology, Department of Radiology, University of Aberdeen

Professor Adrian Grant, Director, Health Services Research Unit, University of Aberdeen

Professor F D Richard Hobbs, Professor of Primary Care & General Practice, Department of Primary Care & General Practice, University of Birmingham

Professor Peter Jones, Head of Department, University Department of Psychiatry, University of Cambridge

Professor Sallie Lamb, Research Professor in Physiotherapy/Co-Director, Interdisciplinary Research Centre in Health, Coventry University

Professor Julian Little, Professor of Epidemiology, Department of Medicine and Therapeutics, University of Aberdeen

Professor Stuart Logan, Director of Health & Social Care Research, The Peninsula Medical School, Universities of Exeter & Plymouth

Professor Tim Peters, Professor of Primary Care Health Services Research, Division of Primary Health Care, University of Bristol

Professor Ian Roberts, Professor of Epidemiology & Public Health, Intervention Research Unit, London School of Hygiene and Tropical Medicine

Professor Peter Sandercock, Professor of Medical Neurology, Department of Clinical Neurosciences, University of Edinburgh

Professor Mark Sculpher, Professor of Health Economics, Centre for Health Economics, Institute for Research in the Social Services, University of York

Professor Martin Severs, Professor in Elderly Health Care, Portsmouth Institute of Medicine

Dr Jonathan Shapiro, Senior Fellow, Health Services Management Centre, Birmingham

Ms Kate Thomas,
Deputy Director,
Medical Care Research Unit,
University of Sheffield

Professor Simon G Thompson, Director, MRC Biostatistics Unit, Institute of Public Health, Cambridge

Ms Sue Ziebland,
Senior Research Fellow,
Cancer Research UK,
University of Oxford

Current and past membership details of all HTA 'committees' are available from the HTA website (www.ncchta.org)

Diagnostic Technologies & Screening Panel

Members

Chair,
Dr Ron Zimmern, Director of the Public Health Genetics Unit, Strangeways Research Laboratories, Cambridge

Ms Norma Armston, Freelance Consumer Advocate, Bolton

Professor Max Bachmann Professor Health Care Interfaces, Department of Health Policy and Practice, University of East Anglia

Professor Rudy Bilous Professor of Clinical Medicine & Consultant Physician, The Academic Centre, South Tees Hospitals NHS Trust

Dr Paul Cockcroft, Consultant Medical Microbiologist/Laboratory Director, Public Health Laboratory, St Mary's Hospital, Portsmouth

Professor Adrian K Dixon, Professor of Radiology, Addenbrooke's Hospital, Cambridge

Dr David Elliman, Consultant in Community Child Health, London

Professor Glyn Elwyn, Primary Medical Care Research Group, Swansea Clinical School, University of Wales Swansea

Dr John Fielding, Consultant Radiologist, Radiology Department, Royal Shrewsbury Hospital

Dr Karen N Foster, Clinical Lecturer, Dept of General Practice & Primary Care, University of Aberdeen

Professor Antony J Franks, Deputy Medical Director, The Leeds Teaching Hospitals NHS Trust

Mr Tam Fry, Honorary Chairman, Child Growth Foundation, London

Dr Edmund Jessop, Medical Adviser, National Specialist Commissioning Advisory Group (NSCAG), Department of Health, London

Dr Jennifer J Kurinczuk, Consultant Clinical Epidemiologist, National Perinatal Epidemiology Unit, Oxford

Dr Susanne M Ludgate, Medical Director, Medical Devices Agency, London

Dr William Rosenberg, Senior Lecturer and Consultant in Medicine, University of Southampton

Dr Susan Schonfield, CPHM Specialised Services Commissioning, Croydon Primary Care Trust

Dr Margaret Somerville, Director of Public Health, Teignbridge Primary Care Trust

Professor Lindsay Wilson Turnbull, Scientific Director, Centre for MR Investigations & YCR Professor of Radiology, University of Hull

Professor Martin J Whittle, Head of Division of Reproductive & Child Health, University of Birmingham

Dr Dennis Wright, Consultant Biochemist & Clinical Director, Pathology & The Kennedy Galton Centre, Northwick Park & St Mark's Hospitals, Harrow

Pharmaceuticals Panel

Members

Chair,
Dr John Reynolds, Clinical Director, Acute General Medicine SDU, Oxford Radcliffe Hospital

Professor Tony Avery, Professor of Primary Health Care, University of Nottingham

Professor Stirling Bryan, Professor of Health Economics, Health Services Management Centre, University of Birmingham

Mr Peter Cardy, Chief Executive, Macmillan Cancer Relief, London

Dr Christopher Cates, GP and Cochrane Editor, Bushey Health Centre

Professor Imti Choonara, Professor in Child Health, University of Nottingham, Derbyshire Children's Hospital

Mr Charles Dobson, Special Projects Adviser, Department of Health

Dr Robin Ferner, Consultant Physician and Director, West Midlands Centre for Adverse Drug Reactions, City Hospital NHS Trust, Birmingham

Dr Karen A Fitzgerald, Pharmaceutical Adviser, Bro Taf Health Authority, Cardiff

Mrs Sharon Hart, Managing Editor, *Drug & Therapeutics Bulletin*, London

Dr Christine Hine, Consultant in Public Health Medicine, Bristol South & West Primary Care Trust

Professor Stan Kaye, Professor of Medical Oncology, Consultant in Medical Oncology/Drug Development, The Royal Marsden Hospital

Ms Barbara Meredith, Project Manager Clinical Guidelines, Patient Involvement Unit, NICE

Dr Frances Rotblat, CPMP Delegate, Medicines Control Agency, London

Professor Jan Scott, Professor of Psychological Treatments, Institute of Psychiatry, University of London

Mrs Katrina Simister, New Products Manager, National Prescribing Centre, Liverpool

Dr Richard Tiner, Medical Director, Association of the British Pharmaceutical Industry

Dr Helen Williams, Consultant Microbiologist, Norfolk & Norwich University Hospital NHS Trust

Therapeutic Procedures Panel

Members

Chair,
Professor Bruce Campbell,
Consultant Vascular and General Surgeon, Royal Devon & Exeter Hospital

Dr Mahmood Adil, Head of Clinical Support & Health Protection, Directorate of Health and Social Care (North), Department of Health, Manchester

Dr Aileen Clarke,
Reader in Health Services Research, Public Health & Policy Research Unit,
Barts & the London School of Medicine & Dentistry,
Institute of Community Health Sciences, Queen Mary,
University of London

Mr Matthew William Cooke,
Senior Clinical Lecturer and Honorary Consultant,
Emergency Department,
University of Warwick, Coventry & Warwickshire NHS Trust,
Division of Health in the Community, Centre for Primary Health Care Studies, Coventry

Dr Carl E Counsell, Senior Lecturer in Neurology,
University of Aberdeen

Dr Keith Dodd, Consultant Paediatrician, Derbyshire Children's Hospital

Professor Gene Feder, Professor of Primary Care R&D, Barts & the London, Queen Mary's School of Medicine and Dentistry, University of London

Professor Paul Gregg,
Professor of Orthopaedic Surgical Science, Department of Orthopaedic Surgery,
South Tees Hospital NHS Trust

Ms Bec Hanley, Freelance Consumer Advocate, Hurstpierpoint

Ms Maryann L. Hardy,
Lecturer,
Division of Radiography,
University of Bradford

Professor Alan Horwich,
Director of Clinical R&D, The Institute of Cancer Research, London

Dr Phillip Leech, Principal Medical Officer for Primary Care, Department of Health, London

Dr Simon de Lusignan,
Senior Lecturer, Primary Care Informatics, Department of Community Health Sciences,
St George's Hospital Medical School, London

Dr Mike McGovern, Senior Medical Officer, Heart Team, Department of Health, London

Professor James Neilson,
Professor of Obstetrics and Gynaecology, Dept of Obstetrics and Gynaecology,
University of Liverpool,
Liverpool Women's Hospital

Dr John C Pounsford,
Consultant Physician, North Bristol NHS Trust

Dr Vimal Sharma,
Consultant Psychiatrist & Hon Snr Lecturer,
Mental Health Resource Centre,
Victoria Central Hospital,
Wirrall

Dr L David Smith, Consultant Cardiologist, Royal Devon & Exeter Hospital

Professor Norman Waugh,
Professor of Public Health,
University of Aberdeen

Expert Advisory Network

Members

Professor Douglas Altman,
Director of CSM & Cancer Research UK Med Stat Gp, Centre for Statistics in Medicine, University of Oxford, Institute of Health Sciences, Headington, Oxford

Professor John Bond,
Director, Centre for Health Services Research, University of Newcastle upon Tyne, School of Population & Health Sciences, Newcastle upon Tyne

Mr Shaun Brogan,
Chief Executive, Ridgeway Primary Care Group, Aylesbury

Mrs Stella Burnside OBE,
Chief Executive, Office of the Chief Executive. Trust Headquarters, Altnagelvin Hospitals Health & Social Services Trust, Altnagelvin Area Hospital, Londonderry

Ms Tracy Bury,
Project Manager, World Confederation for Physical Therapy, London

Mr John A Cairns,
Professor of Health Economics, Health Economics Research Unit, University of Aberdeen

Professor Iain T Cameron,
Professor of Obstetrics and Gynaecology and Head of the School of Medicine, University of Southampton

Dr Christine Clark,
Medical Writer & Consultant Pharmacist, Rossendale

Professor Collette Mary Clifford,
Professor of Nursing & Head of Research, School of Health Sciences, University of Birmingham, Edgbaston, Birmingham

Professor Barry Cookson,
Director, Laboratory of Healthcare Associated Infection, Health Protection Agency, London

Professor Howard Stephen Cuckle,
Professor of Reproductive Epidemiology, Department of Paediatrics, Obstetrics & Gynaecology, University of Leeds

Professor Nicky Cullum,
Director of Centre for Evidence Based Nursing, University of York

Dr Katherine Darton,
Information Unit, MIND – The Mental Health Charity, London

Professor Carol Dezateux,
Professor of Paediatric Epidemiology, London

Mr John Dunning,
Consultant Cardiothoracic Surgeon, Cardiothoracic Surgical Unit, Papworth Hospital NHS Trust, Cambridge

Mr Jonothan Earnshaw,
Consultant Vascular Surgeon, Gloucestershire Royal Hospital, Gloucester

Professor Martin Eccles,
Professor of Clinical Effectiveness, Centre for Health Services Research, University of Newcastle upon Tyne

Professor Pam Enderby,
Professor of Community Rehabilitation, Institute of General Practice and Primary Care, University of Sheffield

Mr Leonard R Fenwick,
Chief Executive, Newcastle upon Tyne Hospitals NHS Trust

Professor David Field,
Professor of Neonatal Medicine, Child Health, The Leicester Royal Infirmary NHS Trust

Mrs Gillian Fletcher,
Antenatal Teacher & Tutor and President, National Childbirth Trust, Henfield

Professor Jayne Franklyn,
Professor of Medicine, Department of Medicine, University of Birmingham, Queen Elizabeth Hospital, Edgbaston, Birmingham

Ms Grace Gibbs,
Deputy Chief Executive, Director for Nursing, Midwifery & Clinical Support Servs, West Middlesex University Hospital, Isleworth

Dr Neville Goodman,
Consultant Anaesthetist, Southmead Hospital, Bristol

Professor Alastair Gray,
Professor of Health Economics, Department of Public Health, University of Oxford

Professor Robert E Hawkins,
CRC Professor and Director of Medical Oncology, Christie CRC Research Centre, Christie Hospital NHS Trust, Manchester

Professor F D Richard Hobbs,
Professor of Primary Care & General Practice, Department of Primary Care & General Practice, University of Birmingham

Professor Allen Hutchinson,
Director of Public Health & Deputy Dean of ScHARR, Department of Public Health, University of Sheffield

Dr Duncan Keeley,
General Practitioner (Dr Burch & Ptnrs), The Health Centre, Thame

Dr Donna Lamping,
Research Degrees Programme Director & Reader in Psychology, Health Services Research Unit, London School of Hygiene and Tropical Medicine, London

Mr George Levvy,
Chief Executive, Motor Neurone Disease Association, Northampton

Professor James Lindesay,
Professor of Psychiatry for the Elderly, University of Leicester, Leicester General Hospital

Professor Rajan Madhok,
Medical Director & Director of Public Health, Directorate of Clinical Strategy & Public Health, North & East Yorkshire & Northern Lincolnshire Health Authority, York

Professor David Mant,
Professor of General Practice, Department of Primary Care, University of Oxford

Professor Alexander Markham,
Director, Molecular Medicine Unit, St James's University Hospital, Leeds

Dr Chris McCall,
General Practitioner, The Hadleigh Practice, Castle Mullen

Professor Alistair McGuire,
Professor of Health Economics, London School of Economics

Dr Peter Moore,
Freelance Science Writer, Ashtead

Dr Andrew Mortimore,
Consultant in Public Health Medicine, Southampton City Primary Care Trust

Dr Sue Moss,
Associate Director, Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton

Professor Jon Nicholl,
Director of Medical Care Research Unit, School of Health and Related Research, University of Sheffield

Mrs Julietta Patnick,
National Co-ordinator, NHS Cancer Screening Programmes, Sheffield

Professor Robert Peveler,
Professor of Liaison Psychiatry, University Mental Health Group, Royal South Hants Hospital, Southampton

Professor Chris Price,
Visiting Chair – Oxford, Clinical Research, Bayer Diagnostics Europe, Cirencester

Ms Marianne Rigge,
Director, College of Health, London

Dr Eamonn Sheridan,
Consultant in Clinical Genetics, Genetics Department, St James's University Hospital, Leeds

Dr Ken Stein,
Senior Clinical Lecturer in Public Health, Director, Peninsula Technology Assessment Group, University of Exeter

Professor Sarah Stewart-Brown,
Director HSRU/Honorary Consultant in PH Medicine, Department of Public Health, University of Oxford

Professor Ala Szczepura,
Professor of Health Service Research, Centre for Health Services Studies, University of Warwick

Dr Ross Taylor,
Senior Lecturer, Department of General Practice and Primary Care, University of Aberdeen

Mrs Joan Webster,
Consumer member, HTA – Expert Advisory Network

Feedback

The HTA Programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.ncchta.org) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@soton.ac.uk
http://www.ncchta.org

ISSN 1366-5278