Health Technology Assessment 2005; Vol. 9: No. 12

A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy

J Dinnes, J Deeks, J Kirby and P Roderick

March 2005

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.

A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy

J Dinnes,[1*] J Deeks,[2] J Kirby[3] and P Roderick[4]

[1] Wessex Institute for Health Research and Development, University of Southampton, UK
[2] Centre for Statistics in Medicine, Oxford, UK
[3] Southampton Health Technology Assessments Centre, University of Southampton, UK
[4] Health Care Research Unit, University of Southampton, UK

* Corresponding author

Declared competing interests of authors: none

Published March 2005

This report should be referenced as follows:

Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Technol Assess* 2005;**9**(12).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE and *Science Citation Index Expanded (SciSearch®)* and *Current Contents®*/Clinical Medicine.

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 02/31/01. 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.

Abstract

A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy

J Dinnes,[1*] J Deeks,[2] J Kirby[3] and P Roderick[4]

[1] Wessex Institute for Health Research and Development, University of Southampton, UK
[2] Centre for Statistics in Medicine, Oxford, UK
[3] Southampton Health Technology Assessments Centre, University of Southampton, UK
[4] Health Care Research Unit, University of Southampton, UK
* Corresponding author

Objectives: To review how heterogeneity has been examined in systematic reviews of diagnostic test accuracy studies.
Data sources: Centre for Reviews and Dissemination's Database of Abstracts of Reviews of Effects (DARE).
Review methods: Systematic reviews that evaluated a diagnostic or screening test by including studies that compared a test with a reference test were identified from DARE. Reviews for which structured abstracts had been written up to December 2002 were screened for inclusion. Data extraction was undertaken using standardised data extraction forms.
Results: A total of 189 systematic reviews met the inclusion criteria. The median number of studies included was 18. Meta-analyses have a higher number with a median of 22 studies compared with 11 for narrative reviews. Graphical plots to demonstrate the spread in study results were provided in 56% of meta-analyses; in 79% these were plots of sensitivity and specificity in the receiver operating characteristic (ROC) space. Statistical tests to identify heterogeneity were used in 32% of reviews: 41% of meta-analyses and 9% of reviews using narrative syntheses. The χ^2 test and Fisher's exact test to assess heterogeneity in individual aspects of test performance were the most common. In contrast, only 16% of meta-analyses used correlation coefficients to test for a threshold effect. A narrative synthesis was used in 30% of reviews. Of the meta-analyses, 52% carried out statistical pooling alone, 18% conducted only summary receiver operator characteristic (SROC) analyses and 30% used both methods of statistical synthesis. For those undertaking SROC analyses, the main differences between the models used were the weights chosen for the regression models, although in 42% of cases the use of, or choice of, weight was not provided. The proportion of reviews using statistical pooling alone has declined from 67% in 1995 to 42% in 2001, with a corresponding increase in the use of SROC methods, from 33% to 58%. However, two-thirds of those using SROC methods also carried out statistical pooling rather than presenting only SROC models. Reviews using SROC analyses also tended to present their results as some combination of sensitivity and specificity rather than using alternative, perhaps less clinically meaningful, means of data presentation such as diagnostic odds ratios. Three-quarters of meta-analyses attempted to investigate statistically possible sources of variation, using subgroup analysis or regression analysis. The impact of clinical or socio-demographic variables was investigated in 74% of these reviews and test- or threshold-related variables in 79%. At least one quality-related variable was investigated in 63% of reviews. Within this subset, the most commonly considered variables were the use of blinding, sample size, the reference test used and the avoidance of verification bias.
Conclusions: The emphasis on pooling individual aspects of diagnostic test performance and the under-use of statistical tests and graphical approaches to identify heterogeneity perhaps reflect the uncertainty in the most appropriate methods to use and also greater familiarity with more traditional indices of test accuracy. This indicates the difficulty and complexity of carrying out such reviews. In these cases it is strongly suggested that meta-analyses are carried out with the involvement of a statistician familiar with the field. Further methodological work on the statistical methods available for combining diagnostic test accuracy studies is needed, as are sufficiently large, prospectively designed primary studies of diagnostic test accuracy comparing two or more tests for the same target disorder. Use of individual patient data meta-analysis in diagnostic test accuracy reviews should be explored to allow heterogeneity to be considered in more detail.

Contents

List of abbreviations

AUC	area under the curve
CI	confidence interval
DARE	Database of Abstracts of Reviews of Effects
DOR	diagnostic odds ratio
FN	false negative
FP	false positive
HSROC	hierarchical summary receiver operating characteristic
IQR	inter-quartile range
LR	likelihood ratio
MRI	magnetic resonance imaging
OR	odds ratio
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RCT	randomised controlled trial
ROC	receiver operating characteristic
RR	relative risk
SROC	summary receiver operating characteristic
STARD	Standards for reporting of diagnostic accuracy
TN	true negative
TP	true positive
UTI	urinary tract infection

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

Systematic reviews of therapeutic interventions are now commonplace in many if not most areas of healthcare, and in recent years interest has turned to applying similar techniques to research evaluating diagnostic tests. One of the key parts of any review is to consider how similar or different the available primary studies are and what impact any differences have on studies' results. Between-study differences or heterogeneity in results can result from chance, from errors in calculating accuracy indices or from true heterogeneity, that is, differences in design, conduct, participants, tests and reference tests. An important additional consideration for diagnostic studies is differences in results due to variations in the chosen threshold for a positive result for either the index or reference test.

Dealing with heterogeneity is particularly challenging for diagnostic test reviews, not least because test accuracy is conventionally represented by a pair of statistics and not by a single measure of effect such as relative risk, and as a result a variety of statistical methods are available that differ in the way in which they tackle the bivariate nature of test accuracy data:

- methods that undertake independent analyses of each aspect of test performance
- methods that further summarise test performance into a single summary statistic
- methods that use statistical models that simultaneously consider both dimensions of test performance.

The validity of a choice of meta-analytical method depends in part on the pattern of variability (heterogeneity) observed in the study results. However, currently there is no empirical guidance to judge which methods are appropriate in which circumstances, and the degree to which different methods yield comparable results. All this adds to the complexity and difficulty of undertaking systematic reviews of diagnostic test accuracy.

Objectives

Our objective was to review how heterogeneity has been examined in systematic reviews of diagnostic test accuracy studies.

Methods

Systematic reviews that evaluated a diagnostic or screening test by including studies that compared a test with a reference test were identified from the Centre for Reviews and Dissemination's Database of Abstracts of Reviews of Effects. Reviews for which structured abstracts had been written up to December 2002 were screened for inclusion. Data extraction was undertaken using standardised data extraction forms by one reviewer and checked by a second.

Results

A total of 189 systematic reviews met our inclusion criteria and were included in the review. The median number of studies included in the reviews was 18 [inter-quartile range (IQR) 20]. Meta-analyses (n = 133) have a higher number with a median of 22 studies (IQR 20) compared with 11 (IQR 13) for narrative reviews (n = 56).

Identification of heterogeneity

Graphical plots to demonstrate the spread in study results were provided in 56% of meta-analyses; in 79% of cases these were in the form of plots of sensitivity and specificity in the receiver operating characteristic (ROC) space (commonly termed 'ROC plots').

Statistical tests to identify heterogeneity were used in 32% of reviews: 41% of meta-analyses and 9% of reviews using narrative syntheses. The χ^2 test and Fisher's exact test to assess heterogeneity in individual aspects of test performance were most commonly used. In contrast, only 16% of meta-analyses used correlation coefficients to test for a threshold effect.

Type of syntheses used

A narrative synthesis was used in 30% of reviews. Of the meta-analyses, 52% carried out statistical pooling alone, 18% conducted only summary receiver operator characteristic (SROC) analyses and 30% used both methods of statistical synthesis. Of the reviews that pooled accuracy indices, most pooled each aspect of test performance separately with only a handful producing single summaries of test performance such as the diagnostic odds ratio. For those undertaking SROC analyses, the main differences between the models used were the weights chosen for the regression models. In fact, in 42% of cases (27/64) the use of, or choice of, weight was not provided by the review authors.

The proportion of reviews using statistical pooling alone has declined over time from 67% in 1995 to 42% in 2001, with a corresponding increase in the use of SROC methods, from 33% to 58%. However, two-thirds of those using SROC methods also carried out statistical pooling rather than presenting only SROC models. Reviews using SROC analyses also tended to present their results as some combination of sensitivity and specificity rather than using alternative, perhaps less clinically meaningful, means of data presentation such as diagnostic odds ratios.

Investigation of heterogeneity sources

Three-quarters of meta-analyses attempted to investigate statistically possible sources of variation, using subgroup analysis (76) or regression analysis (44). The median number of variables investigated was four, ranging from one variable in 20% of reviews to over six in 27% of reviews. The ratio of median number of variables to median number of studies was 1:6.

The impact of clinical or socio-demographic variables was investigated in 74% of these reviews and test- or threshold-related variables in 79%. At least one quality-related variable was investigated in 63% of reviews. Within this subset, the most commonly considered variables were the use of blinding (41% of reviews), sample size (33%), the reference test used (28%) and the avoidance of verification bias (25%).

Conclusions

The emphasis on pooling individual aspects of diagnostic test performance and the under-use of statistical tests and graphical approaches to identify heterogeneity perhaps reflect the uncertainty in the most appropriate methods to use and also greater familiarity with more traditional indices of test accuracy. This is an indication of the level of difficulty and complexity of carrying out these reviews. It is strongly suggested that in such reviews meta-analyses are carried out with the involvement of a statistician familiar with the field.

Recommendations for further research

The following areas are suggested for further research.

- Further methodological work on the statistical methods available for combining diagnostic test accuracy studies is needed.
- Sufficiently large, prospectively designed primary studies of diagnostic test accuracy that compare two or more tests for the same target disorder are needed so that sources of heterogeneity are minimised and comparative accuracy can be established in a wide spectrum of patients.
- Use of individual patient data meta-analysis in diagnostic test accuracy reviews should be explored to allow heterogeneity to be considered in more detail.

Chapter 1
Introduction

Diagnosis is a fundamental element of patient care. It can sometimes be established by clinical examination or history taken alone, but it usually depends on additional laboratory, radiology or pathology tests. The diagnostic process is important for establishing the presence of specific disorders, for informing or monitoring patient prognosis and therapy and in reassuring clinicians and/or patients where the disorders are ruled out. Although there is increasing interest in the evaluation of diagnostic tests and strategies in terms of their impact on patient management and outcomes, there are practical difficulties in designing studies to evaluate these outcomes. The majority of studies focus on estimating diagnostic test accuracy, i.e. the results of one (or more) tests for the detection of a given disorder are compared with the results of some reference standard for that disorder, in a group of patients suspected of having the target disorder to produce a variety of indices of test accuracy.

Systematic reviews of diagnostic test accuracy

Systematic reviews provide a means of synthesising information from a number of studies to "establish where the effects of healthcare are consistent and research results can be applied across populations, settings, and differences in treatment; and where effects may vary significantly".[1] Systematic reviews of therapeutic interventions are now commonplace in many if not most areas of healthcare, and in recent years interest has turned to applying similar techniques to research evaluating diagnostic tests. The HTA Programme has funded a large number of such reviews, and the Cochrane Collaboration has also decided to develop a new database for reviews of diagnostic test accuracy to be incorporated in the Cochrane Library.

Systematic reviews of any form of intervention follow key stages, including formulation of the question, setting of inclusion criteria, searching the literature, quality assessment and data extraction of included studies and synthesis of the evidence. Work is ongoing to develop each of these stages specifically for diagnostic test reviews, for example in literature searching[2] and quality assessment,[3] and several authors have published general guidelines for the conduct of reviews of test accuracy.[4,5] Meta-analytic techniques for combining diagnostic studies, in order to improve estimation of test accuracy, are also being developed and improved.[4–11] The use of statistical methods to combine test accuracy studies is particularly challenging, not least because test accuracy is conventionally represented by a *pair* of statistics (most often sensitivity and specificity; see Appendix 1) and not by a single measure of effect such as the odds ratio (OR) or relative risk (RR). An introduction to the statistical methods that may be used is provided in Chapter 2. However, before (and during) any study synthesis, it is important to consider how similar or different the available primary studies are.

Sources of heterogeneity in diagnostic test reviews

There is almost always considerable variation between the results of diagnostic studies, possibly to a greater extent than is seen for therapeutic interventions, although this comparison needs confirmation in empirical studies. This may be due at least partially to the fact that the importance of rigorous design has been less well appreciated than for therapeutic interventions, and consequently diagnostic studies have often been retrospective and not conducted according to standard protocols. Between-study differences or heterogeneity in results can result from chance, from errors in calculating accuracy indices or from true heterogeneity,[12] that is, differences in design, conduct, participants, interventions, tests and reference tests.[13,14] There is heterogeneity that arises from biases in the conduct of such studies that can be significantly reduced by rigorous design and the heterogeneity that arises from true differences in the accuracy between different test populations and variation in the test under study. An important additional consideration for diagnostic studies is differences in results due to variations in the threshold for a positive result.

In randomised trials, the statistical outcomes that are considered are usually relative comparisons (such as RRs and ORs) or absolute comparisons

(such as risk differences and differences between means) of event rates between treated and control groups made within each trial. While often there is substantial variation in the event rates in the treated groups and in the placebo group between the trials (as displayed in a L'Abbé plot), there may be little variability in relative or absolute comparisons between these event rates. In contrast, for analyses of diagnostic test accuracy the focus is on the event rates in the diseased (test sensitivity) and in the non-diseased (test specificity), and not on relative or absolute comparisons between diseased and non-diseased groups within studies. Hence the level of heterogeneity observed in test accuracy reviews may be higher than that observed in randomised trials owing to the statistical focus not being on comparisons within studies but on absolute estimates of event rates.

Study design and quality considerations

The existence of bias in diagnostic test research has been recognised for many years, with several authors highlighting the potential influence of various forms of bias relating to the study population, the selection and execution of the tests, interpretation of the tests and the data analysis and presentation,[4–11] some of which are discussed below. Empirical evidence for the impact of many of these quality features on test accuracy is still limited. Two studies[15,16] found several features that significantly over- or underestimated test accuracy, including the use of case–control design with healthy controls and severe cases of disease, use of different reference tests, selective inclusion of patients and retrospective data collection.[16]

Verification bias

Verification bias occurs where the decision to undertake or apply the reference test is influenced by the result of the experimental test[9,17,18] (also called **ascertainment bias** or **work-up bias**). There are two potential elements to verification bias:

1. Partial verification occurs where only a subgroup of patients who received the index test undergo the reference test (e.g. where the reference test is unpleasant or invasive, such as biopsy or angiography). This incomplete verification may be equal in test positive and test negative cases (i.e. cases missing at random), or it may be differential where those most likely to have the disease tend to undergo the reference test.
2. Differential verification occurs where different reference tests are used and can occur under two scenarios:
   (a) where those most likely to have the disease tend to undergo the reference test, thereby overestimating sensitivity and underestimating specificity, or
   (b) where different tests are used according to the results of the experimental test (e.g. index test positive patients may undergo a more invasive and probably more accurate reference test than those who tested negative on the index test).

For example, in a study of radionuclide ventriculography for detecting coronary artery disease, 31% of index test positive cases underwent verification compared with only 14% of index test negative cases.[19] The better the test under evaluation, or at least the stronger the investigator's faith in the test, the greater will be the tendency to verify preferentially index test positives and the greater will be the bias introduced.[20]

Use of an appropriate reference test

Standard techniques for assessing diagnostic tests assume that a definitive reference test is available, that is, that the reference test used is as close to 100% accurate as can be. However, it may be either that the available test is far from perfect or that such a test simply does not exist. For example, the diagnosis of metastatic liver cancer can never be definitively determined even at autopsy. The key issue really is not to find a test that confirms a textbook definition of disease but to find a test that has practical consequences for patient management, hence the use of the term 'target disorder' as opposed to 'disease'.

In some contexts where a single definitive reference test is unavailable, a reference **strategy** may be used, where the reference diagnosis is made on the basis of clinical information in combination with a battery of other tests.[9] **Incorporation** bias occurs where the experimental test is used as part of the reference strategy, that is, the experimental test and reference tests are not independent, leading to overestimation of both sensitivity and specificity.[21]

Even the most definitive reference test may have considerable inaccuracies, for example, microbiological studies of sputum for the detection of tuberculosis often fail to detect mycobacteria that may be picked up by nucleic acid amplification tests, and will incorrectly classify patients with tuberculosis as false-positive results.[20] Walter and colleagues[22] refer to a 'substantial body of literature' demonstrating that reference tests may frequently be imperfect. Serious inaccuracies

in the reference test will lead to over- or underestimation of the true accuracy of a new test. If the index and reference test are conditionally independent, then the new test's characteristics will be underestimated (non-differential misclassification); if the two tests are perfectly correlated, or if the new test makes the same errors as the reference test, the accuracy of the new test will be overestimated,[9] potentially appearing perfectly accurate regardless of its association with true disease status.[18]

Blinding

The interpretation of many diagnostic tests involves some degree of subjective interpretation. In clinical practice, test interpretation can be influenced by both the knowledge of the results of other tests and by the specific clinical characteristics of the person being tested. **Diagnostic review bias** occurs where knowledge of the reference test result influences interpretation of the index test, whereas **test review bias** refers to the opposite situation. **Clinical review bias** is said to occur where knowledge of patients' clinical characteristics or other test results influences test interpretation (index or reference test). For example, to evaluate adequately the accuracy of ultrasound for the detection of rotator cuff tear, observers should not have access to the results of other imaging tests such as X-ray or magnetic resonance imaging (MRI). This should be distinguished from observer variability, which will occur in interpretation of almost any test.

The recommended solution to these biases is to perform a 'blinded' study, where both tests are interpreted without knowledge of the clinical characteristics or the test results[17] to ensure that it is only the diagnostic contribution of the test itself that is being evaluated. Of course, this is not the same as routine clinical practice where prior information is used to evaluate the results of subsequent tests. Blinding is particularly important where a new test is intended to replace an existing test, for example, the use of MRI instead of ultrasound for the assessment of shoulder pain. Where clinical factors play a significant role in assisting test interpretation, such as in the shoulder pain example above, or where a new test is intended to supplement an existing test, it may be more appropriate to identify the additional diagnostic value added by the test, rather than essentially evaluating the test in isolation.

Study design

Cohort studies assemble patients at risk for a disease in whom both the new test and the reference test are performed, whereas case–control studies assemble patients with the disease and controls without the disease (on the basis of the reference test results) and compare the index test results in the two groups.[21] Case–control studies tend to be at higher risk from bias: cases tend to be selected on the basis of a positive reference test result and the result of the test under evaluation ascertained after true disease status is known; the prevalence of the target disorder tends to be higher than in cohort studies (or than in practice); and cases and controls are often selected from opposite ends of the disease spectrum, e.g. severe cases and healthy controls.[23] The 'best' cohort studies are prospective in design, with consecutive recruitment of patients; this allows evaluation on the full spectrum (see below) presenting in that setting, the collection of appropriate baseline information and implementation of rigorous protocols for testing.

Clinical heterogeneity

There is some limited evidence that test accuracy statistics may not be generalisable from diagnostic test studies to patients in clinical practice as a result of variations in case mix of participants. The term 'spectrum' (referring to the range of pathological, clinical and co-morbid patient or disease characteristics) was first introduced in 1978 by Ransohoff and Feinstein[24] as an explanation for why many initially promising diagnostic tests are later 'rejected as worthless'. Since then, 'spectrum bias' has been used to describe scenarios where the accuracy indices obtained in one study cannot be assumed to apply to other patients in other contexts and also where test accuracy has been seen to vary according to subgroups of patients within the same study. Such characteristics can be likened to effect modifiers in therapeutic interventions.

It is often assumed that indices of test accuracy such as sensitivity and specificity are fixed (for any given threshold) and that varied is the predictive value between groups with different disease prevalence, the effect of which is easy to estimate. However, theoretical examples[9,21,25] indicate that where spectrum bias is present, either sensitivity or specificity would be expected to change. Sensitivity would be expected to increase where test results become more extreme in patients with the most severe disease (i.e. more likely to test positive). Specificity is affected by the range of alternative diagnoses in those without the target disorder that could cause false positive (FP) results. The range of such diagnoses is likely to be wider in studies that have a lower prevalence of

the target disorder. Variations in case mix, therefore, may affect the generalisability of a study's accuracy results. A classic example is exercise testing for coronary heart disease where the extent of disease (number of diseased arteries) and clinical presentation (e.g. presence or absence of angina) influence the likelihood of an abnormal result and the characteristics of 'non-diseased' (such as gender) affects likelihood of an FP result.[26] Urine dipstick results for urinary tract infection (UTI) are affected by age (as the type of bacteria causing UTI change with age), by pregnancy or by the presence of concomitant medical conditions which affect the response to infection (leukaemia) or type of bacteria (prostatic obstruction).

Variation in test(s)

Just as variations in the timing, duration and dosage or intensity of a therapeutic intervention can affect effectiveness, diagnostic test accuracy may be affected by variations in timing, in technical aspects of any equipment or materials used and, inter- and intra-observer and laboratory variations. Similar variations in the reference standards used must also be considered.

Threshold effects

A source of heterogeneity that is unique to meta-analyses of diagnostic tests is variations in the cut-off chosen to indicate test positivity. Statistics used to report the results of diagnostic tests (e.g. sensitivity and specificity) by nature present a test result as binary, i.e. a test is either positive or negative, disease either present or absent. However, in practice, a test result could be used to predict a good or bad outcome or to differentiate mild from severe disease. The majority of tests effectively produce continuous data such that an arbitrary cut-off point (diagnostic threshold) is applied to define positive and negative test outcomes. In some cases, such as laboratory tests, this could be explicit numerical cut-offs. Imaging tests, such as mammograms, can be interpreted on a categorical scale ranging from definitely normal to definitely abnormal, with various categories of suspicion in between. These thresholds can also be affected by variation between laboratories or between observers[9] – one observer's 'mildly abnormal' may be another's 'definitely abnormal'. The diagnostic classification of patients therefore depends on whether the measurement of a given trait is above or below some defined cut-off or threshold value,[27] and the threshold chosen may vary between studies of the same test. The higher the cut-off value chosen, the higher the specificity and lower the sensitivity estimates will be. The issue of threshold effects is discussed further in Chapter 2.

Chapter 2

Approaches to analysis of heterogeneity in meta-analyses of test accuracy

Introduction

Published systematic reviews of diagnostic test accuracy use a variety of statistical approaches to both meta-analysis and investigations of heterogeneity. This section reviews the most commonly used approaches for meta-analysis and within each approach considers the options for identification and investigation of heterogeneity, explaining the basic statistical methodology. The section concludes with a discussion on appropriate selection of methods.

The approach followed for both meta-analysis and detecting and investigating heterogeneity depends on the summary statistics that are selected for analysis. The options for summarising diagnostic accuracy in an individual study focus on evaluating either the performance of the test in diseased and in non-diseased individuals (test sensitivity and specificity), or the implications of positive and negative test results [positive and negative likelihood ratios (LRs) and predictive values]. Whichever perspective is taken, meta-analysis of diagnostic accuracy is complicated by there being two dimensions to diagnostic performance that require separate estimation.

Approaches to meta-analysis for diagnostic test accuracy can be grouped into three categories according to the way in which they tackle the bivariate nature of test accuracy data:

- methods that undertake independent analyses of each aspect of test performance
- methods that further summarise test performance into a single summary statistic
- methods that use statistical models that simultaneously consider both dimensions of test performance.

Methods of meta-analysis for combining trials are usually categorised as either fixed effect or random effects methods. Fixed effect methods estimate an average effect, assuming that the variability between studies is explicable by sampling variability. Random effects methods explicitly estimate the variability between studies in addition to the average effect. Where the variability between studies is explicable by chance, the random effects estimate of variability will be zero, and the results are exactly the same (or very close to) the results of a fixed effect method.[28]

For systematic reviews of diagnostic test accuracy, there are often high levels of variability between studies which cannot be explained by chance. Thus methods which estimate (or in some way take account of) the extra variability, such as random effects methods, are particularly important.

Graphical plots for displaying heterogeneity

'ROC plots'

Receiver operating characteristic (ROC) curves allow the authors of primary studies evaluating a diagnostic test to display a full picture of that test's accuracy, that is, the relevant combinations of sensitivity and specificity for different thresholds for positivity can be read from the curve similar to that depicted in *Box 1*, rather than presenting only a single sensitivity and specificity pair. Systematic reviewers can also use ROC space to plot the various pairs of sensitivities and specificities from the primary studies to display heterogeneity in the studies' results. These plots are commonly termed 'ROC plots'. Some divergence of the study results around a central point is to be expected by chance, but variations in other factors, such as patient selection and features of study design, may increase the observed variability.[15]

There is also one important extra source of heterogeneity to consider: variation introduced by changes in diagnostic threshold, as discussed in the section 'Threshold effects' (p. 4). Unlike other sources of variability, variation of the diagnostic threshold introduces a particular pattern into the ROC plot of study results, such that the points will demonstrate curvature as depicted in the ROC curve in *Box 1*.

Threshold-like patterns of variability will also be created by changes in the population from which

BOX 1 *Receiver operating characteristic (ROC) curves*

ROC curves are used in studies of diagnostic accuracy to depict the pattern of sensitivities and specificities observed when the performance of the test is evaluated at several different diagnostic thresholds. *Figure 1* is a ROC curve from a study of the detection of endometrial cancer by endovaginal ultrasound.[30] Women with endometrial cancer are likely to have increased endometrial thicknesses: very few women who do not have cancer will have thicknesses exceeding a high threshold, whereas very few women with endometrial cancer will have thicknesses below a low threshold. This pattern of results is seen in *Figure 1*, with the 5-mm threshold demonstrating high sensitivity (0.98) and poor specificity (0.59), whereas the 25-mm threshold demonstrates poor sensitivity (0.24) but high specificity (0.98).

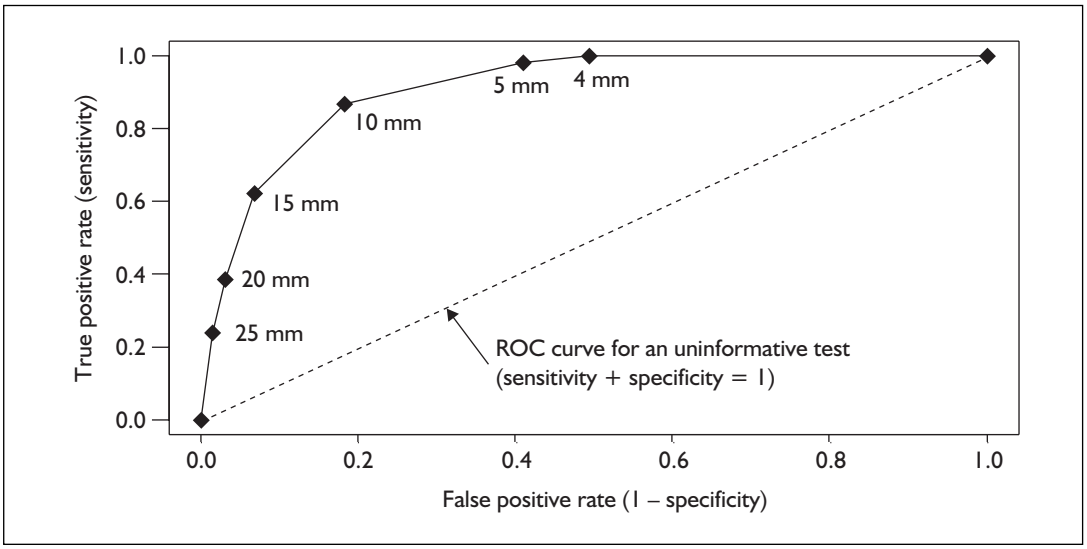


FIGURE 1 *ROC plot of endovaginal ultrasound for detecting endometrial cancer*

The overall diagnostic performance of a test can be judged by the position of the ROC line. Poor tests have ROC lines close to the rising diagonal, whereas the ROC lines for perfect tests would rise steeply and pass close to the top left-hand corner, where both the sensitivity and specificity are unity. ROC plots are used in systematic reviews to display the results of a set of studies, the sensitivity and specificity from each study being plotted as a separate point in the ROC space.

the diseased and non-diseased samples are drawn, without there being any explicit change in threshold. If the measured value of a diagnostic marker increases or decreases depending on other factors in addition to the presence of disease (such as age), then the plot of points from different studies may demonstrate a ROC-type curve. This type of variability has been termed 'implicit threshold effects'.[29]

The existence of threshold effects complicates the statistical analysis. Where such effects are in action, the values of sensitivities and specificities are negatively correlated and cannot be treated as independent quantities in any analysis.

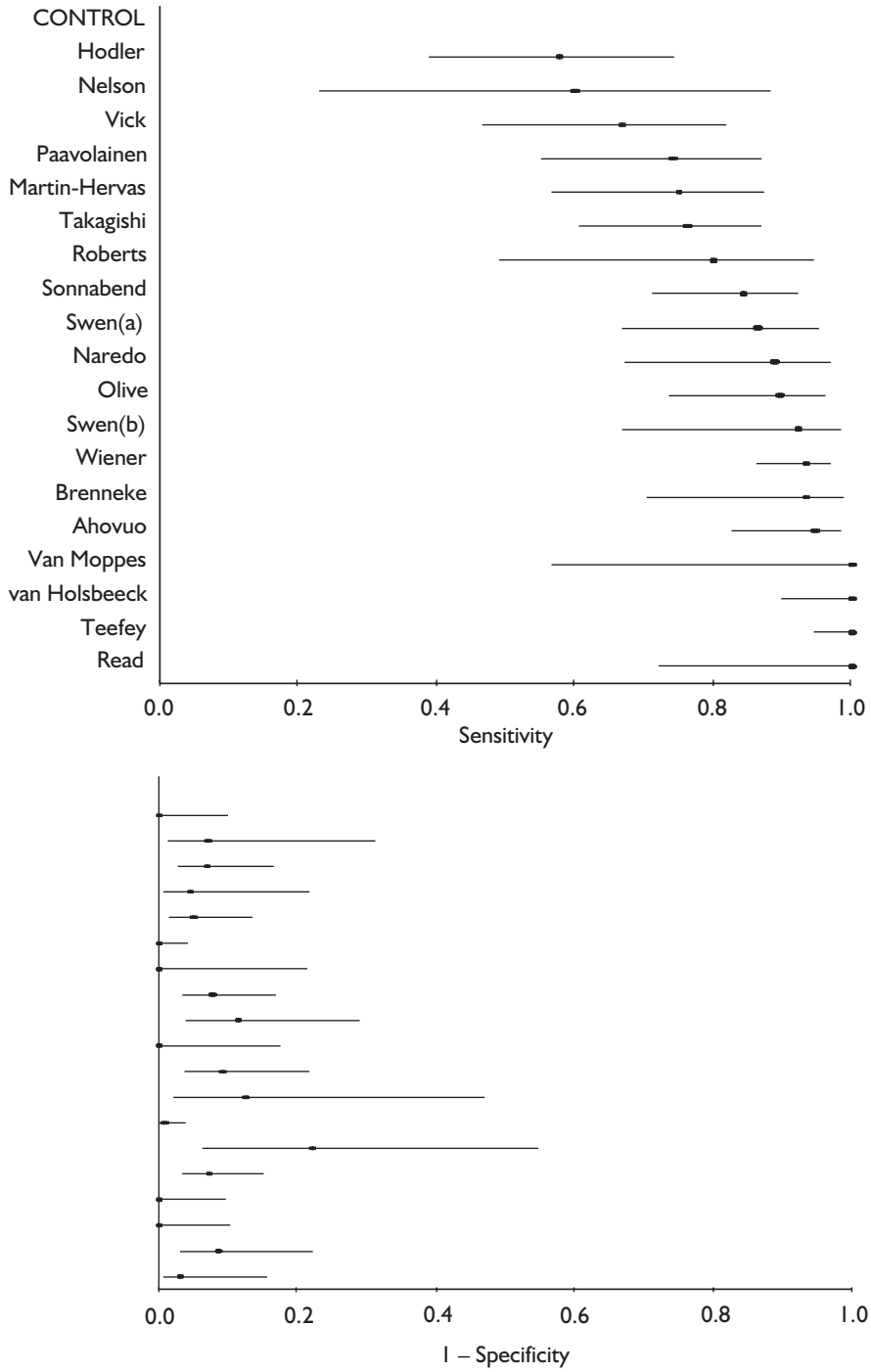
Alternative graphical plots

Accuracy indices and their respective confidence intervals can be presented on Forest plots as for sensitivity and 1 – specificity in *Box 2*. These plots clearly demonstrate the variation in accuracy between studies – homogeneity can be assessed by the extent of overlap in the confidence intervals (CIs). However, it is not possible to detect any pattern in the pairs of sensitivity and specificity, for example due to threshold effects. LRs, predictive values or diagnostic ORs can also be plotted in the same way, but again this will only show the overall heterogeneity (or otherwise) of the accuracy index in question.

Another type of plot that occasionally crops up is the '*D* versus *S*' plot or Littenberg–Moses plot.[31] This is a plot of the logarithm of the diagnostic OR, *D*, against the measure of diagnostic threshold, *S*, for each study. Estimation of *D* and *S* is the preliminary step when using the summary receiver operating characteristic (SROC) regression method proposed by Moses and colleagues[32] [see the section 'Investigating heterogeneity in DORs with threshold' (p. 11) for further details].

An alternative but rarely used option is the Galbraith plot,[33] the use of which in diagnostic test reviews has recently been described by Lijmer and colleagues.[12] The log OR of each study

BOX 2 *Forest plot of sensitivities and 1 – specificities*



divided by its standard error is plotted against the reciprocal of the standard error so that small studies with less precise results are placed on the left-hand side of the diagram and larger studies plotted on the right. The overall log ORs are represented by a regression line through the origin with two lines either side representing its 95% boundaries. In the absence of heterogeneity, most studies would be expected to lie within the 95% boundaries; those lying near or outside of the boundaries can be examined more closely to identify any obvious differences. As Lijmer and colleagues point out, however, one should bear in mind that such examinations would be *post hoc*.[12]

Meta-analyses of sensitivities and specificities, predictive values and likelihood ratios

Meta-analysis of sensitivities and specificities

Sensitivities and specificities are proportions, and can each be pooled by computing weighted averages in either fixed effect or random effects frameworks.[9]

Weights may be computed using the normal variance equation for a proportion:

$$\text{Weight} = \frac{1}{\text{Var}(p)} = \frac{1}{\text{SE}(p)^2} = \frac{n}{p(1-p)}$$

or by first taking *logit* transformations, and calculating weights from the variance of the *logits*:

$$\text{Weight} = \frac{1}{\text{Var}\left[\log\left(\frac{p}{1-p}\right)\right]} = \frac{1}{\frac{1}{pn} + \frac{1}{(1-p)n}}$$

or simply using the sample size, n, as a weight. Standard inverse variance (fixed effects) and DerSimonian and Laird (random effects) software routines can be used to pool if inverse variance weights are used.[34,35] Where computation of estimates and standard errors are problematic, i.e. when proportions are either zero or one, a zero cell correction of 0.5 can be added to the numbers test positive and test negative in both diseased and non-diseased groups.

An alternative approximate fixed effect approach which avoids using zero cell corrections estimates the overall proportion as

$$p = \frac{\Sigma y_i}{\Sigma n_i}$$

where Σy_i is the sum of all true positives (TPs) (for sensitivity) or true negatives (TNs) (for specificity), and Σn_i is the sum of diseased (for sensitivity) or not diseased (for specificity).[9] The large sample approximation for the standard error of this estimate is

$$\text{SE}(p) = \sqrt{\frac{p(1-p)}{\Sigma n_i}}$$

A second fixed effect alternative involves using maximum likelihood methods in a logistic regression model to provide an estimate of the average effect. The standard logistic regression model is equivalent to a fixed effect model, by which heterogeneity can be accounted for by rescaling the standard errors according to the degree of overdispersion computed from the deviance statistic.[36] This random effects model differs from other models in that the random effect acts in a multiplicative rather than additive manner.

Detecting and investigating heterogeneity in sensitivities and specificities

Heterogeneity in proportions (sensitivity or specificity) can be assessed using a χ^2 test on $k - 1$ degrees of freedom (or alternatively Fisher's exact test when cell counts are small and there are few enough studies to make the computations feasible).[9,37] If a logistic regression model is used to estimate average effects, the deviance statistic can be used to assess heterogeneity (goodness-of-fit).

Investigating potential sources of heterogeneity can be undertaken in a number of ways.

1. Univariate two-group t-tests or Mann–Whitney U-tests can be used to make comparisons between the estimates of sensitivities or specificities between two groups. The t-test compares mean values for the two groups and assumes roughly normal distributions of values; the Mann–Whitney U-test compares the positions of two distributions, assuming similar distributional shapes. Neither approach takes account of differences in weights allocated to the studies to reflect differences in precision.
2. Computing meta-analytical estimates for two subgroups and comparing the estimates using a z-test may be a more powerful approach where studies differ in precision, as the estimates will account for study weights. The z-value is computed as

$$z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{[\text{SE}(\hat{\theta}_1)]^2 + [\text{SE}(\hat{\theta}_2)]^2}}$$

   where $\hat{\theta}_k$, are estimates of the overall effect within each group and $\text{SE}(\hat{\theta}_k)$ are the standard errors of these estimates.[28]
3. An alternative test, which can be used regardless of the number of subgroups, involves explicitly partitioning the overall heterogeneity into that which can be explained by differences between subgroups and that which remains unexplained within the subgroups. If the χ^2 value for the overall unstratified analysis is Q_T

and the χ^2 values for each subgroup are Q_k, the heterogeneity explained by differences between subgroups, Q_B, is given by

$$Q_B = Q_T - \sum_k Q_k$$

which can be compared with critical values of the χ^2 distribution with $k - 1$ degrees of freedom.[28]

4. Meta-regression uses a weighted regression technique to estimate associations between predictor variables and estimates of sensitivity or specificity. When the comparison is between two groups, meta-regression produces similar answers to the z-test approach. However, it also allows investigation of trends in estimates with values of a continuous covariate and investigation of multiple sources of heterogeneity simultaneously. Random effects meta-regression models are usually preferred as they make allowance for residual heterogeneity when assessing statistical significance.[28]
5. The logistic regression model can naturally be extended to include terms for covariates, both categorical and continuous. Unexplained variability can be allowed for by rescaling standard errors, as described above.

Meta-analysis and investigating heterogeneity in predictive values

Predictive values, such as sensitivity and specificity, are proportions, and can be pooled using exactly the same methods. Usually predictive values are not used as a summary statistic, as they depend on both the diagnostic accuracy of the test and the prevalence of disease in the study sample. Where some studies in a meta-analysis have separately recruited diseased and non-diseased participants (as in a case–control study design), they are particularly unhelpful as the observed prevalence in the study sample is determined by the sampling proportion such that the predictive values in the studies may bear no resemblance to the predictive value in clinical practice.

Meta-analyses of positive and negative likelihood ratios

LRs are ratios of probabilities, and in a meta-analysis can be treated as risk ratios [albeit calculated between the columns of a 2×2 table and not the rows as for randomised controlled trials (RCTs)]. A fixed effect weighted average of each LR can be computed using the standard Mantel–Haenszel or inverse variance methods of meta-analysis of risk ratios, and a random effects estimate can be computed using the DerSimonian and Laird method for pooling risk ratios. All these analyses combine risk ratios having applied a log-transform.

Detecting and investigating heterogeneity in likelihood ratios

The heterogeneity of each LR can be tested by standard meta-analysis tests of heterogeneity after combining the statistics in a meta-analysis. Cochran's Q is given by

$$Q = \Sigma w_i(\hat{\theta}_i - \hat{\theta})^2$$

where $\hat{\theta}_i$ is the estimate of the log LR for each study, w_i is the weight given to that study in the meta-analysis and $\hat{\theta}$ is the corresponding meta-analytical summary. For a formal test of homogeneity, the statistic Q will follow a χ^2 distribution on $k - 1$ degrees of freedom under the null hypothesis that the true treatment effect is the same for all trials.

Methods have also been developed which assess the impact of heterogeneity on the results of a meta-analysis. For example, the I^2 statistic is estimated by

$$I^2 = \frac{(Q - df)}{Q} \times 100\%$$

where Q is the χ^2 statistic and df is its degrees of freedom.[14,38] This describes the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error (chance). A value greater than 50% may be considered substantial heterogeneity.

Of the methods used to investigate heterogeneity in sensitivities and specificities, the first four (univariate tests, z-tests, χ^2 tests and meta-regression) can also be used to investigate heterogeneity in LRs.

Computing single statistical summaries of test performance

Three related single statistic summaries of test performance are used in meta-analyses of diagnostic test accuracy: diagnostic ORs, diagnostic effectiveness scores and SROC curves.

Diagnostic odds ratios

Sensitivities and specificities, and positive and negative LRs, can be combined into the same single summary of diagnostic performance, known as the diagnostic odds ratio (DOR). This statistic is

not easy to apply in clinical practice (it describes the ratio of the odds of a positive test result in a patient with disease compared with a patient without disease), but it is a convenient measure to use when combining studies in a systematic review as it is often reasonably constant regardless of variation in diagnostic threshold. The DOR is defined as

$$\text{DOR} = \frac{TP \times TN}{FP \times FN}$$

where *TP*, *TN*, *FP* and *FN* are the numbers of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) diagnoses. It is necessary to add a small quantity (typically 0.5) to all four counts if any of them are zero before computing this statistic to avoid computational problems. Some authors advise doing this routinely in all studies.

The DOR can also be computed from the sensitivity and specificity or from the LRs as

$$\text{DOR} = \frac{\left(\dfrac{\text{sensitivity}}{1 - \text{sensitivity}}\right)}{\left(\dfrac{1 - \text{specificity}}{\text{specificity}}\right)} = \frac{\text{LR +ve}}{\text{LR –ve}}$$

where LR +ve is the LR for a positive result and LR –ve is the LR for a negative result. Note that when a test provides no diagnostic evidence (sensitivity + specificity = 1), the DOR is 1. Considering DORs that correspond to commonly cited guidelines for LRs of 0.2 and 5 and of 0.1 and 10 for convincing and strong diagnostic evidence[39] gives a gauge to values of the DOR which could be usefully high. A DOR of 25 could, for example, correspond to a positive LR of 5 and negative LR of 0.2, whereas a DOR of 100 may correspond to a positive LR of 10 and a negative LR of 0.1, if both criteria are met in the same test.

DORs can be pooled using methods for meta-analysis of ORs commonly used for trials. These include the inverse variance and Mantel–Haenszel fixed effect methods, and the DerSimonian and Laird random effects method. However, the Peto OR is rarely appropriate, as it is biased when there are unequal numbers of diseased and non-diseased cases, and when the OR differs from one.[9]

Effectiveness scores

The effectiveness score is defined as[40]

$$\text{diagnostic effectiveness score} = \frac{\sqrt{3}}{\pi}\left\{\log\left(\frac{\text{sensitivity}}{1 - \text{sensitivity}}\right) - \log\left(\frac{1 - \text{specificity}}{\text{specificity}}\right)\right\}$$

The score quantifies the degree of overlap between the distributions of diseased and non-diseased cases, and the value can be interpreted directly as the number of standard deviations separating the means of the two curves. This interpretation is dependent on the distributions having logistic shapes with equal variance. Hasselblad[40] gives some guidance to interpreting values of the effectiveness score. A test with a value of ≤ 1 does not effectively distinguish between groups, a score of 1 meaning that 27% of diseased women have values 'equivalent' (*sic*, at or below the mean value) to non-diseased women. A value of 3 indicates a test where the overlap is only 3% of the sample. Effectiveness scores are directly linked to DORs through the transformation

$$\text{diagnostic effectiveness score} = \frac{\sqrt{3}}{\pi}\log \text{DOR}$$

Pooling methods for diagnostic effectiveness scores are based on the inverse variance approach, weighting each study by the inverse of the variance of the effectiveness score. As diagnostic effectiveness scores are a simple re-expression of the DOR, the same methods for investigating heterogeneity apply.

DORs, effectiveness scores and SROC curves

If there is any evidence that the diagnostic threshold varies between the studies, the best statistical summary of the results of the studies will be a SROC curve rather than a single sensitivity/specificity point. The full method for deciding on the best fitting summary ROC is explained below, but first it is worth noting that when it can be assumed that the curve is symmetrical around the 'sensitivity = specificity' line and the underlying distributions have a bilogistic form with equal variances [see Investigating heterogeneity in DORs with threshold (p. 11) for more details], an estimate of the best fitting ROC curve can be obtained by pooling DORs (or effectiveness scores). Once the summary DOR has been calculated, the equation of the corresponding ROC curve is given by

$$\text{sensitivity} = \frac{1}{1 + \dfrac{1}{\text{DOR} \times \left(\dfrac{1 - \text{specificity}}{\text{specificity}}\right)}}$$

Methods of testing whether the data can be summarised using a symmetrical ROC curve are described below.

Area under the ROC curve

Where a primary study reports a ROC curve rather than a single point estimate of sensitivity and specificity, meta-analysis may attempt to pool a measure called the area under the curve (AUC). Perfect tests have AUCs of close to 1, whereas poor tests have AUCs close to 0.5.[41] However, ROC curves of different shapes can have the same AUC, so it is not possible to interpret the AUC in terms of a set of unique combinations of sensitivity and specificity unless the shape of the ROC curve is known. If it can be assumed that the data come from a bilogistic distribution with equal variance, then the ROC will have a particular symmetric shape consistent with all points having the same DOR, and there is an algebraic relationship between the AUC and the DOR given by[42]

$$\text{AUC} = \frac{\text{DOR}}{(\text{DOR} - 1)^2}\left[(\text{DOR} - 1) - \ln(\text{DOR})\right]$$

Otherwise estimates and standard errors of the area under the ROC curve can be obtained from each study using trapezoid methods and averaged using a standard inverse variance technique.

Detecting and investigating heterogeneity in diagnostic odds ratios

Heterogeneity in DORs can be tested using Cochran's Q as described in the section 'Detecting and investigating heterogeneity in likelihood ratios' (p. 9). Breslow and Day propose an alternative test of the homogeneity of ORs, based on a comparison of the observed number of events in the intervention groups of each trial (a_i) with those expected when the common treatment effect $\hat{OR}$ is applied (calculation of these expected values involves solving quadratic expressions). The test statistic is given by

$$Q_{BD} = \Sigma\left(\frac{a_i - E[a_i|\hat{OR}]}{v_i}\right)$$

where each trial's variance v_i is computed using the fitted cell counts:

$$v_i = \frac{1}{E[a_i|\hat{OR}]} + \frac{1}{E[b_i|\hat{OR}]} + \frac{1}{E[c_i|\hat{OR}]} + \frac{1}{E[d_i|\hat{OR}]}$$

Under the null hypothesis of homogeneity, Q_{BD} also has a χ^2 distribution on $k - 1$ degrees of freedom.[28]

Investigating heterogeneity in DORs with threshold

Heterogeneity in DORs at different thresholds arises when the diseased and non-diseased groups differ in both the average value of the underlying diagnostic marker and also in the variance of the values. For example, diseased patients may have higher values of a diagnostic marker than non-diseased patients, but also the values for diseased people may be more variable than the values for non-diseased people. Where this is the case, dichotomising the diagnostic marker scale at different points will yield different DORs. The ORs at higher cutpoints will be higher than those at lower cutpoints. When the points are plotted as a ROC curve, the curve will not be symmetric about the sensitivity = specificity line. The values of sensitivity at high values of specificity will be higher than the values of specificity at correspondingly high values of sensitivity.

Asymmetric ROC curves occur when the DOR changes with diagnostic threshold. Moses and colleagues proposed a method for fitting a whole family of SROC curves which allow for a trend in DOR with threshold.[31,32] The method considers the relationship between the DOR and a summary measure of diagnostic threshold, given by the product of the odds of TP and the odds of FP results. As a diagnostic threshold decreases, the numbers of positive diagnoses (both correct and incorrect) increases and the measure of threshold increases.

In the equations and figures which follow, the logarithm of the DOR is denoted by D and the logarithm of the measure of threshold by S. D and S can be calculated using any of the equivalent equations:

$$S = \ln\left(\frac{TPR}{1 - TPR} \times \frac{FPR}{1 - FPR}\right)$$

$$= \text{logit}(TPR) + \text{logit}(FPR)$$

$$D = \ln(\text{DOR}) = \ln\left(\frac{TPR}{1-TPR} \times \frac{1-FPR}{FPR}\right)$$

$$= \ln\left(\frac{\text{LR} + \text{ve}}{\text{LR} - \text{ve}}\right)$$

$$= \text{logit}(TPR) - \text{logit}(FPR)$$

where the logit indicates the log of the odds, as used in logistic regression.

Moses and colleagues' method first considers a plot of the log of the DOR (D) against the measure of threshold (S) calculated for each of the studies. They then propose computing the best fitting straight line through the points on the graph. If the equation of the fitted line is given by

$$D = a + bS$$

testing the significance of the estimate of the slope parameter b tests whether there is significant variation in diagnostic performance with threshold. If the line can be assumed horizontal (i.e. $b = 0$), the DOR does not change with threshold and the method yields symmetrical ROC curves, similar to those obtained from directly pooling ORs as explained above. However, if there is a significant trend in the DOR with diagnostic threshold (i.e. $b \neq 0$), then the ROC curves are asymmetric, the summary ROC curve being calculated as

$$\text{sensitivity} = \frac{1}{1 + \dfrac{1}{e^{a/(1-b)} \times \left(\dfrac{1-\text{specificity}}{\text{specificity}}\right)^{(1+b)/(1-b)}}}$$

Estimates of the parameters a and b can be obtained from either ordinary least-squares regression (which weights each study equally), weighted least-squares regression (where the weights can be taken as the inverse variance weights of the diagnostic log OR, or simply the sample size) or robust methods of regression (which are not so strongly influenced by outliers). Although weighting by inverse variance carries appeal in that it combines studies according to the precision of their estimates of the OR, it is problematic when sensitivity or specificity (and hence ORs) is high, as the equation for the approximate variance of a log DOR becomes biased when any of the counts of TPs, TNs, FPs or FNs is close to zero.[43] Some authors describe the equally weighted regression as being a random effects model.[7] This is not technically true, as the model does not estimate the variance of the observed effects (which is the basis of random effects analyses), but reflects the observation that random effects analyses do give studies more equal weightings than fixed effect analyses.

Expositions of Moses and colleagues' method commonly formulate it in terms of the sum and differences in the *logits* of the TP and FP rates. As shown in the equations above, log DOR is in fact the difference of these *logits*, whereas logarithm of the measure of diagnostic threshold is the sum of these *logits*. Hence the choice of notation: D for the difference and S for the sum.

It has been suggested that the validity of the method can be improved by only including points which have values of sensitivity and specificity within a clinically meaningful range and omitting points with sensitivities and specificities below some stated threshold.[10] This is likely to lead to overestimation of the diagnostic accuracy as it systematically excludes the poorest estimates of test accuracy.

Investigating heterogeneity in DOR with other factors

Heterogeneity in DORs for potential sources of heterogeneity can be investigated by the first four methods mentioned in the section 'Detecting and investigating heterogeneity in sensitivities and specificities' (p. 8): univariate tests between subgroups, z-test based comparisons of estimates between groups, χ^2 tests and meta-regression. The use of these methods without allowing for threshold effects makes an assumption that variation in threshold does not affect DORs. This is equivalent to assuming that any underlying ROC curve is symmetrical.

Investigating heterogeneity in DOR with thresholds and other factors

If it is important to allow for variation of DOR with threshold at the same time as investigating other sources of heterogeneity, then Moses and colleagues' model can be extended to allow for covariates.[6,7] A covariate, X, can be added to the regression equation for each potential effect modifier:

$$D = a + bS + c_1X_1$$

The exponential of each of these terms estimates multiplicative increases in DORs (relative ORs) for each factor. An underlying assumption of these

models is that the shape of the SROC curves is not affected by covariates.

A further extension to the model allows for different shapes for the ROC curves indicated by the covariates. To do this, interaction terms between covariates and thresholds are included in the model:

$$D = a + bS + c_1X_1 + d_1SX_1$$

If the covariate indicates, say, differences between two tests, this model is equivalent to fitting separate SROC curves for each test. A problem with this model is that it becomes difficult to judge the importance of differences between the curves, as they may differ both in average diagnostic accuracy and shape, and possibly cross-over. Commonly, points on the curves denoted by Q^* are identified for each subgroup and compared. Q^* is the value on the curve at which sensitivity is equal to specificity, and is the point where the threshold parameter, S, is equal to zero.[32] It is estimated by first computing the DOR when the threshold parameter is zero as DOR = exp(a), where a is the intercept value estimated from the regressions equation and inserting it into the equation

$$Q^* = \frac{\sqrt{\text{DOR}}}{1 + \sqrt{\text{DOR}}}$$

The value of Q^* may not be particularly useful when the range of estimates of sensitivity and specificity from the studies does not include values near the Q^* point. A comparison between Q^* values in this situation would be a comparison between two extrapolated points, and is unlikely to be reliable.

Joint statistical modelling of sensitivity and specificity

Moses and colleagues' SROC method has limitations in that it does not provide an estimate of the average sensitivity/specificity operating point, does not properly weight study estimates and is based on a regression model where the explanatory variable is measured with error. Hence it does not provide appropriate estimates for standard errors for statistical inference for sources of heterogeneity.

Recently, several methods based on generalised linear models have been developed which aim to improve on the SROC approach. The sophistication of these methods differs in their abilities to (1) account for the correlation between sensitivity and specificity due to threshold effects, (2) account for the differences in precision between study estimates, (3) estimate the variability in parameter estimates which is not explicable by chance (i.e. estimate random effects) and (4) report results as clinically meaningful parameters.

Extension of the logistic regression model by Mol and colleagues

Mol and colleagues investigated heterogeneity jointly in sensitivity and specificity by using a logistic regression model in which each study contributed two binomial samples – one of the diseased group (TPs out of sample size for diseased) and one for the non-diseased group (FPs out of sample size of non-diseased).[44] An indicator variable was included in the model for disease group. The parameter estimate for this indicator variable was interpreted as the log DOR. Additional covariates were included representing study characteristics and were investigated as sources of heterogeneity.

One limitation of Mol and colleagues' model is that it was not stratified by study and therefore could not account for the correlation within studies of sensitivities and specificities due to threshold effects.

Extension of the logistic regression model

A simple extension of the logistic regression model outlined in the section 'Detecting and investigating heterogeneity in sensitivities and specificities' (p. 8) involves fitting two separate logistic regression models, one for sensitivities and one for specificities, and accounting for threshold variation in the analysis of sensitivity by including the estimate of specificity (maybe grouped into several categories) as an explanatory variable in the model of sensitivity, and vice versa. Covariates for sources of heterogeneity can be added to each model and feasibly have different effects on sensitivity and specificity. The degree to which these models appropriately deal with threshold effects is unclear, as there will be measurement error in the sensitivities and specificities which will not be accounted for when they are used as explanatory variables, and there is no specification of the shape of the underlying ROC curve.

The Rutter and Gatsonis hierarchical summary receiver operating characteristic (HSROC) mode

The Rutter and Gatsonis HSROC model can be conceived as an extension of Moses and

colleagues' model.[45,46] A hierarchical (or multilevel or mixed) model allows for uncertainty at different levels. For this application, two levels are considered: variation first within studies and second between studies. Multilevel models have been used elsewhere for meta-analysis. Rutter and Gatsonis developed their hierarchical model based on the latent scale logistic regression model formulation. Binomial data are entered for each study, i, as in the model by Mol and colleagues, with the diseased and non-diseased states, j, coded as 0.5 and –0.5, respectively. A variable is included in the model to indicate study. The following non-linear regression model is fitted:

$$\text{logit}(\pi_{ij}) = (\theta_i + \alpha_i\, dis_{ij})\exp(-\beta\, dis_{ij})$$

where π_{ij} is the proportion test positive. The model yields parameter estimates for θ_i (the threshold parameter), α_i (log DOR) and β, which allows for asymmetry in the underlying ROC curve. θ_i and α_i are usually fitted as random effects, so that their average value and variation across studies are estimated. The shape parameter β can only be estimated as a fixed effect. The original publications reported results for models fitted using a Bayesian approach in WinBUGS (which requires specification of prior distributions for all model parameters). Recently, an empirical Bayes' version of the model has been described which can be fitted in PROC NLMIXED in SAS.[29]

The model produces estimates of the mean and variance of the log DOR, the mean and variance of the threshold parameter and an estimate of the shape parameter. Estimates and CIs for the average operating point, expressed as either sensitivity/specificity or positive and negative LRs, can be obtained by combining these estimates.

An SROC curve can be constructed by computing values of sensitivity across the range of specificities using the following equation:

$$\text{sensitivity} = \frac{1}{1 + \exp\left[-\hat{\alpha}\exp(-0.5\hat{\beta}) - \ln\left(\frac{1-\text{specificity}}{\text{specificity}}\right)\exp(-\hat{\beta})\right]}$$

The model can be extended by including covariates for potential sources of heterogeneity. Covariates can be added to accuracy, threshold and shape components of the model, and are usually fitted as fixed effects. The significance of covariates can be evaluated by testing the model terms for the covariates, and differences may be noted in whether covariates alter (a) DORs, (b) the threshold and (c) the shape of the ROC curve.

Fitting HSROC models uses complex iterative mathematical algorithms which are occasionally sensitive to starting values and sometimes do not converge.

The bivariate normal model

The bivariate normal model uses similar hierarchical models such as the Rutter and Gatsonis method, but preserves the sensitivity/specificity parameterisation of the studies, rather than converting test values to estimates of DORs.[47] The mean values and variances of *logit* transformations of sensitivity and specificity are estimated, as is a correlation between sensitivity and specificity acknowledging the pairing of data within each study and the possibility of threshold effects. The model assumes that the *logit* sensitivities and *logit* specificities have normal distributions, and uses the asymptotic variance estimates to compute study weights.[48,49] The model can be fitted in PROC MIXED in SAS, and recently a variation of the model allowing for the binomial errors has been proposed which can be fitted in PROC NLMIXED.

The model provides an estimate of the average operating point, together with 95% CIs (or a bivariate confidence region). Sources of heterogeneity can be investigated by adding covariates to the model. The effect of each covariate on sensitivity and specificity is estimated separately. Although the model does not directly estimate DORs, thresholds and SROC curves, it is possible to transform the parameters to obtain these estimates.

Fitting bivariate normal models also uses complex iterative mathematical algorithms which are occasionally sensitive to starting values and do not always converge.

Selecting a method of meta-analysis

The validity of a choice of meta-analytical method depends in part on the pattern of variability (heterogeneity) observed in the study results. If there is no heterogeneity between the studies, the best summary estimate of test performance will be a single point on the ROC curve (the operating

point), and there will be no sources of heterogeneity to investigate. When there is heterogeneity, there is usually still interest in estimating the average operating point, but also in knowing how the studies vary around the average operating point and how the operating point changes with potential sources of heterogeneity.

If a threshold effect exists, values of sensitivity and specificity (and likewise positive and negative LRs) are not independent. It has been demonstrated that when there is a threshold effect, separate pooling of sensitivity and specificity (or LRs) that ignores the correlation between them may underestimate the average sensitivity and specificity, whereas fitting a ROC curve will provide a more appropriate estimate.[32] Where there is no heterogeneity, both averaging sensitivity and specificity and ROC-based methods give similar results.

In some circumstances, threshold variation may lead to variability in one dimension of test accuracy but not the other.[9] For example, sensitivity may be high in all studies and specificity variable, or vice versa. These are also circumstances in which both ROC curve methods and independent assessments of sensitivity and specificity give similar estimates of the operating point.

Usually where there is threshold variation, approaches to analysis that assume independence of sensitivity and specificity are unlikely to be appropriate. Approaches to meta-analysis either: compute alternative statistics that are invariant to threshold; attempt to estimate the average (or summary) ROC curve; or use statistical models that properly account for the correlation. The latter are preferred.

These observations have led some analysts to propose decision processes based on testing for heterogeneity and threshold effects as a way of deciding which method of analysis should be used.[37] For example, it has been suggested that if tests of heterogeneity of sensitivities and specificities are non-significant, separate pooling of sensitivities and specificities is appropriate. In a similar vein, the correlation between sensitivity and specificity can be tested using a rank correlation test, such as Spearman's rho, and threshold methods used only when the correlation is high or its magnitude large.

These approaches are criticised on two counts. First, tests of heterogeneity and correlation may have inadequate power to detect the effects they are investigating,[10] and second, multistage statistical processes where a method is chosen on the results of a previous significance test can occasionally give incorrect *P* values.

Some analysts argue that as there is an underlying ROC curve for every test that describes the pattern of test performance at different thresholds, all meta-analyses should estimate ROC curves, regardless of whether empirical evidence of a threshold effect exists. The methods described in the sections 'The Rutter and Gatsonis hierarchical summary receiver operating characteristic (HSROC) mode' (p. 13) and 'The bivariate normal model' (p. 14) will produce a point estimate for the operating point with zero estimates of variability when there is no heterogeneity, in addition to being able to deal with situations where there are multiple sources of heterogeneity. Other analysts prefer to use the simplest method possible, and only add complexity when absolutely needed. This approach is usually preferred by medical journals, which prefer not to confuse their readers with unnecessarily complicated statistical methods.

However, there is no empirical guidance to judge which methods are appropriate in particular circumstances and the degree to which different methods yield comparable results. Until empirical investigations have been undertaken, selection of a method must be based on a judgement of whether the mathematical properties of a particular approach are likely to match the scenario to which they are applied.

Notably, comparisons between methods which have been published have been made on the basis of comparing the ability of the alternative methods to estimate an average operating point (sensitivity and specificity values), and not their ability to detect correctly sources of heterogeneity.

Chapter 3
Methods

Aim of the review

The aim of the project was to review how heterogeneity has been examined in systematic reviews of diagnostic test studies.

Eligibility criteria

To be included, reviews must have evaluated a diagnostic or screening test by including studies that compared a test with a reference test. Reviews of studies using a randomised control design that does not allow the calculation of test accuracy were excluded. Studies were assessed for inclusion by one reviewer.

Literature search

The Centre for Reviews and Dissemination's Database of Abstracts of Reviews of Effects (DARE) was used to identify existing systematic reviews of diagnostic studies. This is a database of quality-assessed systematic reviews identified by handsearching key major medical journals, by regular searching of bibliographic databases and by scanning grey literature since 1994 (further details about DARE can be found at http://agatha.york.ac.uk/darehp.htm).

Diagnostic reviews indexed in DARE up to April 2001 had already been screened to identify diagnostic reviews for a previously funded HTA project[3] and were automatically included in this one. Those reviews indexed between April 2001 and December 2002 were also screened for inclusion. Only those for which structured abstracts had been written were considered eligible.

Further searches of primary electronic databases were not undertaken, as we did not consider the systematic identification of all diagnostic test reviews to be necessary for a methodological review. However, additional systematic reviews not indexed on DARE but meeting our inclusion criteria that were identified *ad hoc* were included.

Data extraction

A data extraction form for recording relevant information from each systematic review was designed and piloted (see Appendix 2). Data were extracted on a variety of items, including:

- the experimental test, reference tests and condition tested for
- the review methodology including the literature search and approach to quality assessment
- review synthesis methods and approach to identifying heterogeneity statistically in study results
- methods of exploration of variability in study results and variables investigated.

The full systematic reviews were prescreened independently by two reviewers. Those meeting the inclusion criteria were data extracted by one reviewer and the completed data extraction forms checked against the full paper by a second reviewer. Any disagreements were resolved by consensus or by referral to a third reviewer if necessary.

Data synthesis

A narrative synthesis is presented. The reviews are considered primarily in terms of the statistical methods used, and the Results section is structured to reflect the steps involved in the synthesis of diagnostic test accuracy studies, i.e.

- identification of heterogeneity via graphical presentation of study results and statistical testing
- meta-analysis
- investigation of sources of heterogeneity.

Chapter 4
Results

Summary of reviews identified

Of 312 identified systematic reviews, 189 met our inclusion criteria and were included in the review. *Figure 2* provides a flowchart of the review selection process. The 124 excluded reviews and reasons for their exclusion are presented in Appendix 3. Summary details of the included reviews, according to whether they used a narrative (n = 56, 30%) or a statistical method of synthesis (n = 133, 70%), are provided in *Tables 1–4*; fuller details of the reviews are given in Appendices 4–6.

Description of review methods

The reviews cover a wide range of target disorders and test types, from the low technology of clinical examination for the detection of diseases such as left-sided heart failure,[50] deep vein thrombosis[51] or carpal tunnel syndrome[52] at one end, to highly equipment-intensive tests such as nucleic acid amplification tests for detecting infection[53–55] or positron emission tomography for the detection of cancer or Alzheimer's disease.[56] Appendix 4 has complete details of the review topics.

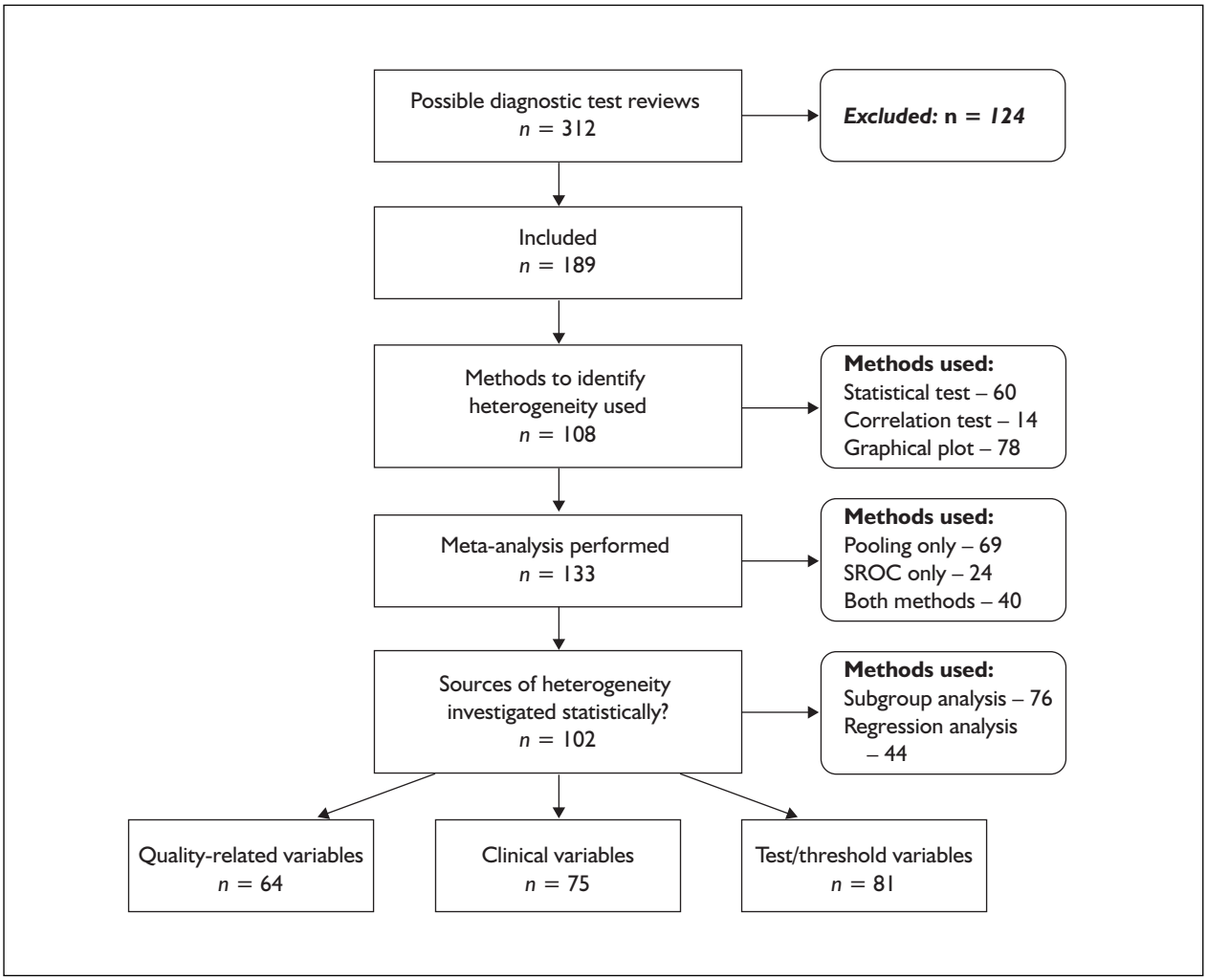


FIGURE 2 *Flowchart of reviews*

TABLE 1 *Summary of reviews found*

		Total *n* (%)	Statistical *n* (%)	Narrative *n* (%)
Total no. of reviews		189	133 (70%)	56 (30%)
Review methods				
MEDLINE only *electronic* source		99 (52%)	78 (59%)	21 (38%)
No. using language restriction	Restricted	111 (59%)	78 (59%)	33 (59%)
	English only	94 (84%)	63 (80%)	31 (94%)
	No restriction	27 (14%)	20 (15%)	7 (13%)
	Not stated	51 (27%)	35 (26%)	16 (29%)
No. using quality restrictions	Restricted	94 (50%)	69 (52%)	25 (45%)
	Appropriate ref. test	71 (76%)	59 (86%)	12 (48%)
	Blinding used	18 (19%)	16 (23%)	2 (8%)
	Prospective only	15 (16%)	9 (13%)	6 (24%)
	Avoidance of verif. bias	14 (15%)	12 (17%)	2 (8%)
	Adequate sample descr.	9 (10%)	5 (7%)	4 (16%)
	Consecutive enrolment	8 (9%)	8 (12%)	0
	Adequate test descr.	2 (2%)	2 (3%)	0
	Complete follow-up	2 (2%)	2 (3%)	0
	No restriction	95 (50%)	64 (48%)	31 (54%)
No. using quality assessment	Not conducted	58 (31%)	40 (30%)	18 (32%)
	Conducted	131 (69%)	93 (70%)	38 (68%)
	Authors' own	88 (67%)	68 (73%)	20 (53%)
	Existing tool	43 (33%)	25 (27%)	18 (47%)
Median (IQR) no. of studies		18 (IQR 20)	22 (IQR 20)	11 (IQR 13)
	No. studies not reported	7 (4%) reviews	3 (2%)	4 (7%)
Median (IQR) no. of patients		3161 (IQR 6815)	4007 (IQR 7553)	1726 (IQR 3619)
	No. patients not reported	68 (36%) reviews	34 (26%)	24 (43%)

IQR, inter-quartile range.

Just over half (52%) of all reviews included searched only one electronic database (MEDLINE) to identify primary studies (*Table 1*). This was less often the case for narrative reviews (38%) compared with those using statistical syntheses (59%). Some 59% of reviews used language restrictions in their searches; in 84% of these this was to restrict studies to English language only. Only 14% (27/188) of reviews applied no language restrictions. These proportions were similar regardless of whether the reviews carried out narrative or statistical syntheses. Half of reviews applied inclusion criteria to restrict studies to those of a higher standard on at least one quality criterion. Most commonly this was to ensure that studies had compared the index test with an appropriate reference standard (86% of meta-analyses and 48% of narrative reviews applying quality-related criteria). The next most commonly used criteria were to ensure that blinding had been used (19%), to include only prospective studies (16%) and to ensure that verification bias had been avoided (15%). Restriction to higher quality studies was more common in meta-analyses than in narrative reviews and meta-analyses were more likely to apply than one quality-related criterion.

Quality assessment of included primary studies was reported to have been carried out in 69% of reviews (*Table 1*), with most (88/131) using a quality assessment tool apparently developed by the authors themselves (only 43 reported using a previously published tool). An analysis of the items included in a sample of these quality assessment tools is provided by Whiting and colleagues.[3]

The median number of studies included in the reviews was 18 (IQR 20). Meta-analyses have a higher number with a median of 22 studies (IQR 20) compared with 11 (IQR 13) for narrative reviews. The number of patients included in the studies was not clearly reported in 36% of all reviews, less so for narrative reviews (not reported in 43%).

Description of statistical methods used

Summary and full details of the statistical methods used in the reviews are presented in *Table 2* and Appendix 5.

Identification of heterogeneity

Graphical plots to identify heterogeneity

Over half (75/133, 56%) of meta-analyses used graphical plots to demonstrate the spread in study results. In 79% (59/75) of cases, study results were plotted in ROC space, 13 reviews plotted sensitivity and/or specificity on Forest plots and three reviews used *D* versus *S* plots.

Only two of the 56 reviews using a narrative synthesis presented study results graphically, all using ROC plots.

Statistical tests to identify heterogeneity

Statistical tests to identify heterogeneity were used in 60 (32%) of reviews (*Table 2*).

Of the 133 reviews using statistical syntheses, 55 (41%) used statistical tests to identify heterogeneity, most (61%) using the chi-squared test and 11% using Fisher's exact test. In 44 reviews (79%), statistically significant heterogeneity was identified. A further five meta-analyses made a narrative statement regarding the presence of heterogeneity. In contrast, only 16% (21/133) of meta-analyses used correlation coefficients to test for a threshold effect, most (14) choosing the Spearman correlation coefficient.

Five of the reviews using a narrative synthesis used statistical tests to identify heterogeneity (*Table 2*), four of which reported that statistically significant

TABLE 2 *Summary of statistical methods used*

		Total *n* (%)	Statistical *n* (%)	Narrative *n* (%)
Total no. of reviews		189	133 (70%)	56 (30%)
Statistical methods used				
Test for heterogeneity reported		60 (32%)	56 (42%)	5 (9%)
(some reviews used more	χ^2	36 (60%)	34 (61%)	3 (60%)
than one test)	Fisher	7 (12%)	6 (11%)	2 (40%)
	Breslow–Day	5 (8%)	4 (7%)	1 (20%)
	Q statistic (ORs)	3 (5%)	3 (5%)	0
	Kardaun–Kardaun	1 (2%)	1 (2%)	0
	Observed vs predicted values	6 (10%)	5 (8%)	1 (20%)
	Miscellaneous tests[a]	5 (8%)	5 (7%)	0
	Test used but not reported	8 (13%)	8 (14%)	0
Test result	Statistically significant	47 (78%)	44 (79%)	3 (60%)
	Not significant	10 (17%)	10 (18%)	0
	Not reported	4 (7%)	3 (5%)	1 (40%)
Correlation test for threshold		21 (11%)	21 (16%)	0
effects	Spearman correlation		14 (67%)	
	Pearson correlation		3 (14%)	
	Kardaun–Kardaun		1 (5%)	
	Test used but not reported		2 (10%)	
Correlation test result	Significant correlation		14 (67%)	
	No correlation		6 (29%)	
	Not reported		1 (5%)	
Study results plotted graphically		77 (41%)	75 (56%)	2 (4%)
	ROC plot	57 (74%)	59 (79%)	2 (100%)
	Forest *Se* and/or *Sp*	12 (16%)	12 (16%)	0
	Forest DOR or log DOR	1 (1%)	1 (1%)	0
	D vs *S* plot	3 (4%)	3 (4%)	0
	Miscellaneous plots[b]	12%)	9 (12%)	0

continued

TABLE 2 *Summary of statistical methods used (cont'd)*

		Total *n* (%)	Statistical *n* (%)	Narrative *n* (%)
Type of synthesis used	*Narrative*		0	56 (100%)
	Pooling methods		109 (82%)	
	Sensitivity/specificity		97 (89%)	NA
	LRs		26 (24%)	NA
	PVs		11 (10%)	NA
	DOR		10 (9%)	
	Effectiveness score		8 (7%)	
	Accuracy		5 (5%)	
	AUC		3 (3%)	
	Miscellaneous[c]		4 (4%)	
	SROC		64 (48%)	
	Weighting not specified		27 (42%)	NA
	Unweighted		13 (20%)	NA
	Inverse variance weighted		11 (17%)	
	Sample size weighted		6 (9%)	NA
	Variance weighted		1 (2%)	NA
	'Weighted'		7 (11%)	NA
	Robust resistant regression		2 (3%)	NA
	Estimated from DOR or ES		3 (5%)	
	Data presentation:			
	DOR		4 (6%)	
	AUC		10 (16%)	
	SROC parameters		7 (11%)	
	*Q**		18[e] (28%)	
	Se or *Sp* at fixed *Sp* or *Se*		20 (31%)	
	SROC curve only presented		10 (16%)	
	Comparison of $\geqslant$ 2 curves		4 (6%)	
	Other methods[d]		2 (3%)	
Paired data considered separately (meta-analyses only)	Yes		12 (9%)	
	No		42 (32%)	
	No paired data (or can't tell)		79 (59%)	
Method of investigating heterogeneity	Not done	17 (9%)	10 (8%)	7 (13%)
	Narrative	68 (36%)	19 (14%)	49 (87%)
	Subgroup	74 (39%)	74 (56%)	NA
	Regression	45 (24%)	45 (34%)	NA
	Method not described	2 (1%)	2 (2%)	NA

[a] Including effectiveness score (2 studies); comparison of fixed vs random effects results (1 study); 'covariate adjustment' (1 study); and goodness of fit test (1 study).
[b] Including: funnel plots using ES (1 study) or log DOR (1 study); scatterplots of AUC (1 study), LR (1 study) or *Se* (1 study) per study; *Se* (1 study) or NPV (1 study) plotted against prevalence; *Se* and *Sp* as function of prevalence (1 study); and *Se*/*Sp* plotted against sample size (1 study).
[c] Including: fraction positive (1 study); correlation coefficient (1 study); Youden index (1 study); odds of FN on index vs reference test (1 study).
[d] Including: ratio of ORs (1 study); estimation of LR, method not reported (1 study).
[e] In two reviews LR was estimated from *Q**.
ES, effectiveness score; NPV, negative predictive value; PV, predictive value; Se, sensitivity; Sp, specificity.

TABLE 3 *Statistical tests and graphical approaches used according to method of synthesis*

Type of synthesis	Narrative 56 (30%)	Statistical syntheses 133 (70%)	Statistical syntheses by method of synthesis used		
			Pooling only 69 (52%)	Pooling and SROC 40 (30%)	SROC only 24 (18%)
Statistical tests used					
Test for threshold effects?	0 (0%)	21 (16%)	5 (7%)	7 (17%)	9 (37%)
Test for heterogeneity?	5 (9%)	55 (41%)	30 (43%)	17 (42%)	8 (33%)
Both tests carried out?	0 (0%)	13 (10%)	4 (6%)	3 (8%)	6 (25%)
Graphical plots presented					
Graphical presentation of results	2 (4%)	76 (57%)	19 (28%)	35 (87%)	22 (92%)
Both statistical test for heterogeneity and graphical plot presented	1 (2%)	29 (22%)	8 (12%)	13 (32%)	8 (33%)
Neither test nor graphical plot	54 (96%)	35 (26%)	28 (41%)	5 (12%)	2 (8%)

heterogeneity was found. A further four reviews specifically stated that the studies were too heterogeneous to be pooled, although no formal evidence for this was provided.

Identification of heterogeneity according to type of synthesis used

Of the 133 (70%) reviews in which meta-analysis was performed, 52% (n = 69) carried out statistical pooling alone, 18% (n = 24) conducted only SROC analyses, and 30% (n = 40) used both methods of statistical synthesis (*Table 3*). Although 57% of meta-analyses presented study results graphically, these were primarily reviews that had used SROC regression models; only 19 (28%) of those using statistical pooling alone presented results graphically.

Tests to identify heterogeneity were used in similar proportions of reviews using pooling alone versus those presenting SROC curves (43% vs 39%); however, tests for threshold effects were more often presented by those using SROC methods (25% compared with 7% of those using pooling alone), although the overall proportion was still low. Many of those using SROC methods stated that these methods allow for the presence of a threshold effect (37/64), so presumably did not see the need to test specifically for threshold effects. Overall, only 10% of meta-analyses carried out both tests to identify general heterogeneity and tests to identify threshold effects, and over one-quarter (26%) neither carried out statistical test nor presented study results graphically.

Type of syntheses used

Meta-analyses of sensitivities and specificities, predictive values and likelihood ratios

Of the 109 reviews that pooled accuracy indices, 89% pooled sensitivity and/or specificity, 24% pooled LRs and 10% pooled predictive values. A further 5% of reviews pooled test 'accuracy', which is the percentage of diagnoses that were correct (i.e. number TP plus number TN as a proportion of all test results).

Pooled single summaries of test performance

Single summaries of test performance, estimated by pooling results from individual studies or by logistic regression methods (akin to fixed-effects pooling) were carried out in only a handful of studies: 9% of those using pooling pooled DORs, 7% pooled the 'effectiveness score' (see Chapter 2), and 3% pooled AUC data from individual studies.

Single summaries of test performance using SROC regression models

For those reviews presenting SROC curves, all except four used regression models such as that described by Moses and colleagues[32] to create the curves. Three of the exceptions estimated SROC curves from the pooled DORs or effectiveness scores and the other did not describe the method used. For the remainder, the main differences between the models used are the weights chosen for the regression model. In 42% of cases (27/64), the use of, or choice of, weight was not provided by the review authors (*Table 2*). In 13 reviews

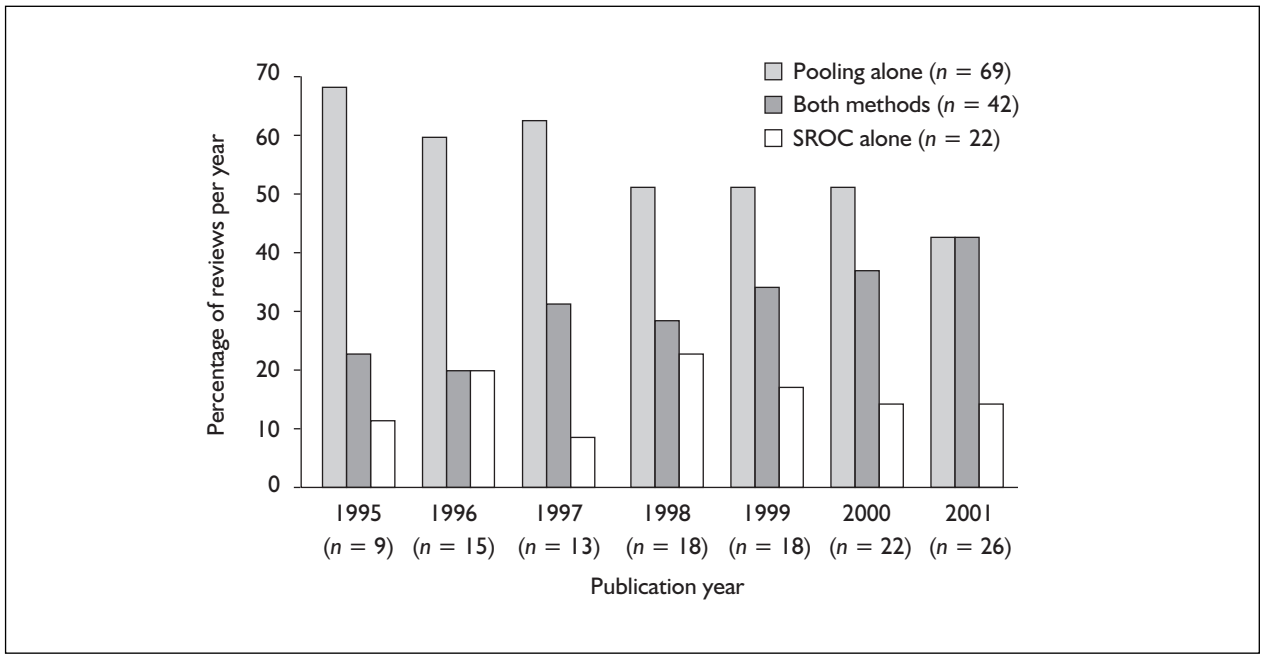


FIGURE 3 *Type of meta-analytic method used by publication year*

(20%), the models were unweighted; in 17%, inverse variance weights were used; and in 9%, sample size weights were used. In a further 11% (6/64), models were simply described as 'weighted'.

As discussed in Chapter 2, SROC curves can be interpreted in several ways. The methods most commonly used in our sample were those that converted certain points of the SROC curve to sensitivity and specificity pairs (*Table 2*): the *Q** (maximum joint sensitivity and specificity) was presented in 28% (18/64) of reviews, sensitivity and specificity pairs were 'read' from the SROC curves in 31% (20/64) of reviews, for example, sensitivity at mean specificity or 95% specificity, or sensitivity and specificity at mean threshold. Ten reviews (16%) chose to provide AUC data and only four (6%) interpreted the SROC curve as a DOR. The underlying SROC model parameters were provided by 11% of reviews, 16% presented the SROC curve only with no summary statistics and 6% compared two or more curves for different tests.

Narrative syntheses of data

A narrative synthesis was used in 56 (30%) of reviews. In eight reviews the authors indicted that this was due to the presence of between-study heterogeneity [see the section 'Identification of heterogeneity (p. 21)], but the remainder did not state whether they had considered using statistical approaches to study synthesis.

Type of statistical synthesis according to publication year

Figure 3 shows the proportion of reviews using each method according to publication year, for those published between 1995 and 2001 (insufficient numbers of reviews were available for other years). The proportion of reviews using statistical pooling alone declined slightly over that period (from 67% in 1995 to 42% in 2001), with a corresponding increase in the use of SROC methods (from 33% of all reviews in 1995 to 58% in 2001). However, two-thirds of those using SROC methods also carried out statistical pooling rather than presenting only SROC models (42/64). The tendency to carry out both methods in the same review has on the whole increased over time.

Data presentation according to type of syntheses used

We hypothesised that where SROC analysis alone was used, reviews would be more likely to present their results as some combination of sensitivity and specificity rather than using alternative, less clinically meaningful, data presentation.

Figure 4 shows an analysis of methods of data presentation in reviews using SROC analysis according to whether or not statistical pooling was also performed. When only SROC analysis was carried out, reviews were more likely to report pairs of sensitivity and specificity data (45% compared with 24% of reviews that also conducted pooling), providing some support for our

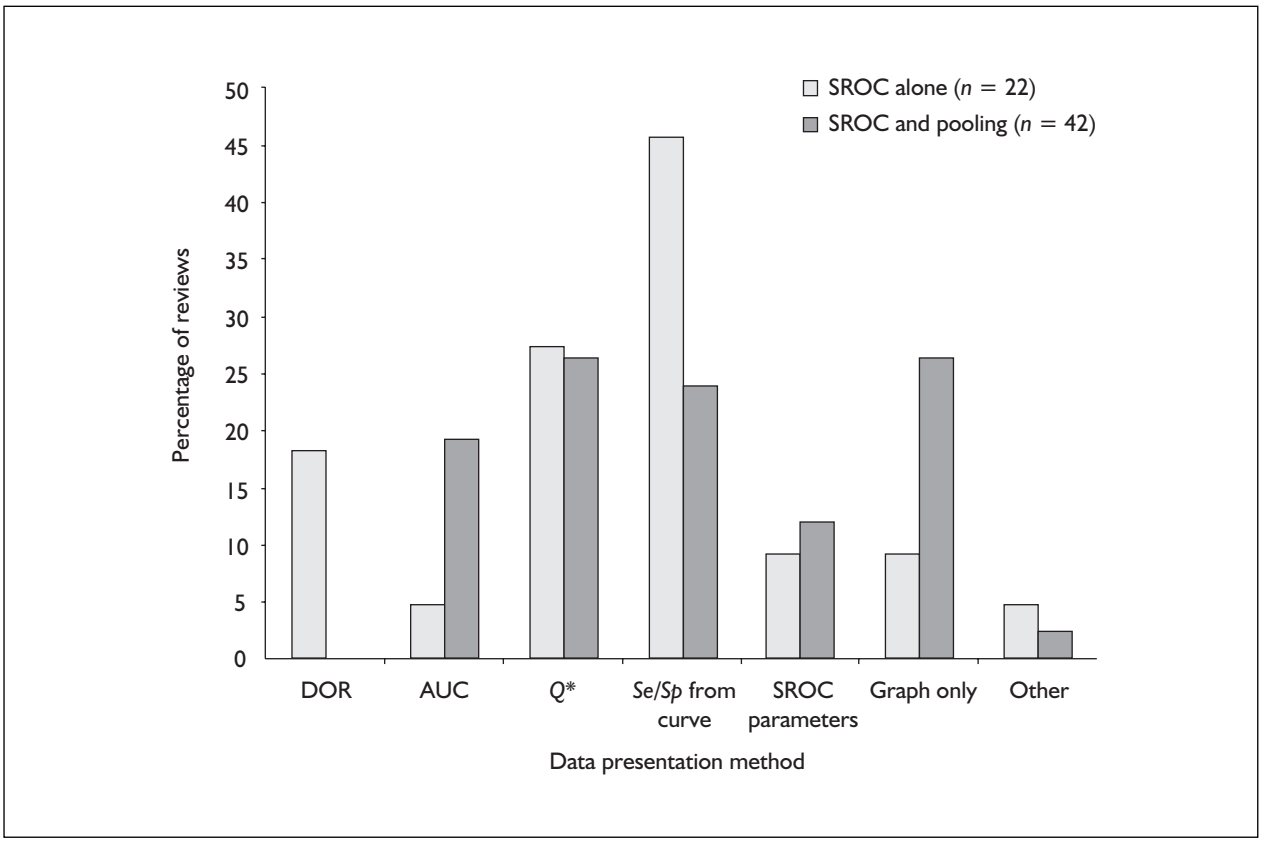


FIGURE 4 *Means of presenting results of SROC analyses (*n *= 64)*

hypothesis. Furthermore, it is not clear whether these sensitivity and specificity pairs were in fact read from the SROC curve or were actually estimated by some form of averaging. It is likely that the point estimate quoted by these reviewers was computed by pooling sensitivities and specificities, and may not have actually been a point on the ROC curve. However, this group of reviews was also more likely to present results as DORs, although actual numbers were small (4/22 reviews). When both pooling and SROC were reported to have been carried out (i.e. where the pooled estimates were clearly presented), reviewers were more likely to present AUC data and were also more likely not to provide a summary statistic to interpret the SROC model, but simply to present the curve.

Consideration of 'paired' data

Although a number of reviews evaluated more than one test, in only 54 of the 133 meta-analyses (41%) were we able to identify primary studies that had evaluated more than one test against a reference standard, and in only 12 of the 54 reviews did the reviewers attempt to deal with the fact that they had 'paired' data, for example by analysing the data from those reviews separately.

Investigation of sources heterogeneity

Methods of investigating heterogeneity

Of the 56 narrative reviews, 49 (87%) carried out a narrative review of factors that might cause variation in the results of the primary studies and seven did not really appear to deal with the question of heterogeneity at all.

Of the meta-analyses, 29 (24%) provided either a narrative discussion of factors affecting heterogeneity (19) or did not consider heterogeneity at all (10). The remaining 102 attempted to investigate statistically possible sources of variation, 74 (56%) using subgroup analysis and 45 (34%) using some form of regression analysis. Regression analyses were usually undertaking by extending the SROC regression model (see Appendix 5), although 10 reviews reported using logistic regression models and one used meta-regression. A further two did not report the method that they had used. For those reviews using subgroup analyses, although several reported *p*-values for the differences between groups, very few reported the test used to detect any statistically significant difference: seven reviews reported using a *t*-test or Mann–Whitney *U*-test to compare subgroups,

TABLE 4 *Statistical investigations of heterogeneity (n = 102)*

		No. (%) of reviews
Median no. of variables considered (IQR)		4 (IQR 4)
% considering only 1 variable		20 (20%)
% considering 2–5 variables		55 (54%)
% considering >6 variables		28 (27%)
Ratio of median no. of variables investigated to median no. of studies included		1:6
Reviews with ratio <1:10		63 (62%)
Categories	**Variables investigated:**	
Quality-related variables	*Not investigated*	38 (37%)
	Investigated	64 (63%)
	Blinding	26 (41%)
	Sample size	21 (33%)
	Ref. test used	18 (28%)
	Verification bias	16 (25%)
	Consecutive enrol	12 (19%)
	Prospective/retrospective	9 (14%)
	Spectrum	6 (9%)
	Disease progression bias	4 (6%)
	Sample description	3 (5%)
	Cohort/case–control design	3 (5%)
	Other QA items	9 (14%)
	Quality 'rating'/score	23 (36%)
Clinical or socio-demographic variables	*Not investigated*	27 (26%)
	Investigated	75 (74%)
	Age	17 (23%)
	Sex	10 (13%)
	Spectrum or clinical-related variables	72 (96%)
Test- or threshold-related variables	*Not investigated*	21 (21%)
	Investigated	81 (79%)
	Test	56 (69%)
	Threshold	31 (38%)
	Publication year	29 (36%)

QA, quality assurance.

two used the χ^2 test and three the Wilcoxon test (paired or unpaired).

Sources of heterogeneity investigated

Table 4 provides a summary of the number and breakdown of variables investigated by the 102 reviews that statistically investigated possible causes of heterogeneity. The median number of variables investigated in these reviews was four, ranging from only one in 20% of reviews to over six in 27% of reviews (*Figure 5*). In general, a large number of variables were investigated in these analyses in comparison with the number of studies included in the review. The ratio of median number of variables to median number of studies was 1:6. Only 38% of reviews complied with the typical recommendation to have at least 10 studies for every characteristic investigated.

At least one study quality-related variable was investigated in 63% (64) of reviews. Within this subset of reviews, the most commonly considered variables were use of blinding (41% of reviews), sample size (33%), the reference test used (28%) and the avoidance of verification bias (25%). The inclusion of an appropriate spectrum of patients and impact of study design chosen were among those considered in a small minority of reviews, 9% and 5%, respectively. Around one-third of reviews (36%) tried to look at the overall effect of study quality on accuracy, for example by classifying studies as low, medium or high quality or by using the quality score to subdivide studies.

The impact of clinical or socio-demographic variables was investigated in 74% of reviews. Most (96%) considered one or more items related to

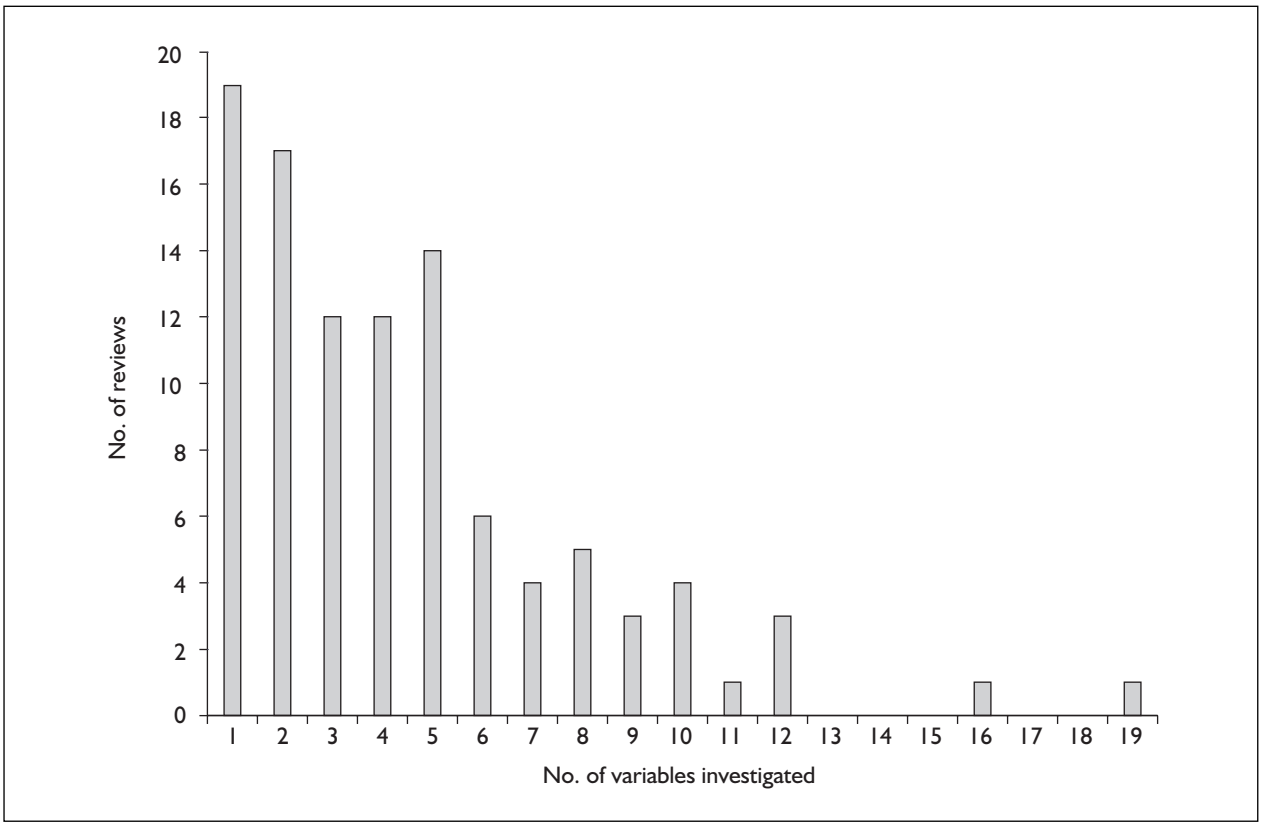


FIGURE 5 *Number of variables investigated per review*

spectrum of disease or disease prevalence, for example stratifying by acute versus elective patients, or referred versus non-referred or by entering covariates such as prevalence of disease or proportion of patients who were symptomatic into the SROC regression model; 23% of reviews considered age and 13% sex.

Test- or threshold-related variables were examined by 79% of the reviews. Most (69%) considered items related to variations in the test used, for example by looking at the effect of variations in the field strength used in MRI, or in the level of expertise of the person interpreting the test. Some 38% of reviews considered threshold by subdividing studies according to threshold used. Publication year, which could be a proxy for changes in a test over time or changes in the patient population tested, was considered important by 36% of reviews.

Chapter 5
Discussion

We found that statistical tests to identify heterogeneity and graphical plots to demonstrate heterogeneity are rarely reported in reviews using narrative syntheses of diagnostic test accuracy and, furthermore, are not always reported in reviews using meta-analytic techniques. One quarter of reviews (35/133) did not report using either technique. Those which did used only statistical tests such as the χ^2 test (26/133; 19%) or graphical approaches to demonstrate heterogeneity (47/133; 35%) rather than using both approaches (29/133; 22%). Graphical presentation of results was mainly carried out by those conducting SROC analysis, that is, individual study results in addition to the SROC curve were presented in ROC space. Of those authors opting only to pool data, less than one-quarter (19/69) used any form of graphical presentation of results, only nine of which presented data on a ROC plot, thereby demonstrating any potential correlation between sensitivity and specificity.

Possible reasons for the low usage of these approaches are unclear, but lack of knowledge of or confidence in the most appropriate methods to use may have contributed to the low use of statistical tests. Tests to detect heterogeneity are known to have low power.[1,12] However, this would not justify the lack of graphical presentation of individual study results. Given the high degree of heterogeneity amongst diagnostic test studies (over three-quarters of those using statistical tests indicated that statistically significant heterogeneity was present), such approaches are a useful aid to conveying complex information, even in reviews choosing to use a narrative synthesis – a perfectly defensible option where studies are highly variable. Plotting pairs of sensitivity and specificity in ROC space is an easy way to display heterogeneity of both indices in addition to allowing potential threshold effects to be detected. It is also true, however, that visual examination of study results to identify heterogeneity also has limited power to detect bias if the number of studies is small. At the very least, reviewers should explicitly acknowledge and assess the potential for heterogeneity to be present, whether statistical approaches to identify it are employed or not.

The wide variation in methods chosen to combine the results of primary studies again perhaps reflects uncertainty in the most appropriate methods to use and also greater familiarity with more traditional indices of test accuracy (e.g. sensitivity and specificity). It would be extremely difficult for us to make a judgement as to whether or not the approach taken by the individual reviewers was appropriate or not without looking at the primary studies, but we have attempted to point out issues that may be of concern.

Narrative reviews may have been carried out owing to assumed but unreported heterogeneity, or to insufficient numbers of studies (the median number of studies in narrative reviews was only half of that in meta-analyses), but few reported having considered the option of using statistical syntheses. Although the median number of studies may have been lower, in principle many did include a sufficient number of studies to consider meta-analysis. Reviewers should recognise that a justification for the approach chosen, whether narrative or statistical, should be provided in systematic reviews.

For those carrying out statistical syntheses, most opted to pool aspects of test performance independently, that is, separate pooling of sensitivity and specificity, positive and negative LRs or predictive values, with little consideration paid to the possibility of a threshold effect. Correlation tests for detecting threshold effects were described in only 16% of reviews, although around half of those using SROC approaches (37/64) stated that they did so because this technique allows for any threshold effect. Of the 69 reviews that only carried out pooling of sensitivity and specificity or LRs (i.e. did not conduct SROC analysis), 39 (57%) did not test for heterogeneity and 64 (93%) did not test for threshold effects. It is likely that the results for a proportion of these studies would differ if methods that allow for heterogeneity and threshold variation were employed.

Reporting of SROC methods is challenging, as the results are not easily interpreted by clinicians. Ideally, a clinician would like to have a point estimate of the sensitivity and specificity of a test,

whereas ROC curves describe a series of estimates. Many authors chose to present the results as some combination of sensitivity and specificity at given points on the SROC curve – the way in which this point was computed was often arbitrary; rarely were a series of potential operating points quoted. No reviews in our sample attempted to pool studies using the advanced statistical methods described in the sections 'The Rutter and Gatsonis hierarchical summary receiver operating characteristic (HSROC) mode' (p. 13) and 'The bivariate normal model' (p. 14), which have been available since 1995. This may be a feature of the age of the articles in our sample, but also may reflect difficulties in applying methods in unconventional software such as WinBUGS, and perhaps lack of publication of cutting edge methodologies in medical journals. These methods offer promise in their ability to estimate properly random effects distributions and for investigating sources of heterogeneity, and should be more commonly used in systematic reviews in years to come.

Regardless of the use of statistical tests to identify heterogeneity, it can certainly be argued that potential sources of heterogeneity should always be investigated in systematic reviews of diagnostic test accuracy studies. However, only three-quarters of meta-analyses in our sample attempted to investigate sources of heterogeneity, and in 20% of those only one characteristic was investigated. However, although such investigations are recommended, they should be limited by the number of studies included in the review. We found that on average one characteristic was investigated for every six studies included in these reviews. This is a possible indication of over-investigation of study characteristics.

The most appropriate choice of variables to be investigated will depend on the specific context of the review and the included studies; however, we have provided an indication of the types of variables that have been chosen. Clinical or socio-demographic variables and test- or threshold-related variables were investigated in 74% and 79% of reviews, respectively, but study design and quality were considered in less than two-thirds of reviews. Blinding, sample size and overall quality classification were the most commonly considered. Half of all reviews in our sample only included studies that met certain quality-related criteria and so may have decided that further investigation of the effect of quality on accuracy was not warranted. However, this is unlikely to be the only explanation. At least some proportion of the lack of investigation of quality-related variables will be due to poor reporting on the part of the authors of the primary studies, and also to the fact no standard quality assessment tool has been available. A new tool for the quality assessment of diagnostic studies, developed using standard scale development techniques, has now been published.[3] The authors hope that in addition to providing a standardised tool for systematic reviewers, the project may also play a role in bringing about greater awareness regarding the important quality issues involved in diagnostic accuracy studies and help to raise the standards of such trials.

Poor reporting is a particular problem with diagnostic accuracy studies such that it is often difficult to ascertain what procedures to avoid bias were actually followed by study authors. The Standards for Reporting of Diagnostic Accuracy (STARD) initiative[57] aims to promote the completeness and quality of reporting of diagnostic accuracy studies similarly to the CONSORT statement for reports of RCTs. Greater awareness of methodological principles for diagnostic accuracy studies will also help inform the design and analysis of primary studies.

We found that nearly all reviews focus on undertaking meta-analyses by comparing the results of a new test with a reference standard. Very few reviews analysed only studies which compared results of several tests in the same patients with a reference standard and only 12/54 (22%) reviews that included at least some 'paired' data on two or more tests considered those studies separately. One can argue that heterogeneity will be less likely to be so problematic in meta-analyses of within-study comparisons between tests, as many of the factors (such as the patient group) will be identical for both tests. Statistical methodology for investigating heterogeneity and threshold effects in studies of paired test comparisons requires further development, but may in time lead to more robust evidence about the relative performance of alternative diagnostic tests.

Other issues highlighted by our review include the significant potential for publication bias in these reviews – 84% restricted studies to those published in English only and 52% searched only one electronic database (MEDLINE). Publication bias is know to be a real problem in reviews of therapeutic interventions.[58,59] Although its extent has not yet been quantified for test accuracy reviews, it seems likely that it will be as much, if not more, of an issue for tests. The retrospective

nature of many diagnostic test studies would imply that authors may only publish if they have found particularly good results with a test.

We have also not been able to study any variation in quality of review methods within different areas of medicine or types of test. This is hard to categorise across reviews and numbers within sub-categories would be small.

A strength of our review was use of the DARE database. Systematic reviews have to meet a certain standard of methodological quality before being included in the database. This does mean that the reviews in our sample are of higher quality than many that are published, so that the situation in practice may be worse than we have demonstrated here. However, it is also notable that the considerable time lag in loading reviews into DARE at the time of our search means that the majority of reviews in our sample were published prior to 2002. Given that comprehensive guidelines on carrying out systematic reviews of diagnostic tests were not published before 2001,[4,9,13,60] it is likely that review methods have improved significantly since that time.

Chapter 6
Conclusions

It is clear that a proportion of published reviews ignore heterogeneity in the analysis and presentation of their results, and simply present average values of sensitivity and specificity (or occasionally LRs). There is a danger that these reviews may be disseminating a misleading message that implies consistency of test performance when in fact the data that they have collected clearly display inconsistency. Such inadequate analyses could in the worst instance lead to inappropriate diagnostic investigations and interpretations and the use of inappropriate interventions.

Where heterogeneity has been considered, the variability in approaches taken is a reflection of the level of difficulty and complexity of carrying out such reviews. The methodology is still developing and there is considerable uncertainty in the most appropriate techniques to use. Recent high-profile guidelines on undertaking diagnostic test reviews[4,9] should go some way to improving standards, as will the Cochrane Collaboration's decision to include diagnostic test accuracy reviews in the Cochrane Library. Nevertheless, carrying out many of the statistical analyses required for these reviews requires a high degree of familiarity with statistics and statistical software packages – there is as yet no truly user-friendly software package that can be used by non-statisticians in the way that packages such as RevMan are used for meta-analyses of therapeutic interventions. It is highly recommended that diagnostic test accuracy meta-analyses should not be carried out without the involvement of a statistician familiar with the field.

Difficulties with investigating heterogeneity at review level also point to the need for sufficiently large, prospective, well-designed, multicentre studies that evaluate a number of diagnostic tests (or variations on a test), in order to establish test accuracy and also allow the investigation of the influence of patient characteristics on accuracy.

Recommendations for future research

The following areas are suggested for further research.

- This review could be updated to identify whether methods of study synthesis have improved since 2001.
- Further methodological work on the statistical methods available for combining diagnostic test accuracy studies is needed to help identify an 'optimal' approach.
- Efficient means of linking diagnostic and therapeutic information are also needed.
- Sufficiently large, prospectively designed primary studies of diagnostic test accuracy that compare two or more tests for the same target disorder are needed so that sources of heterogeneity are minimised and comparative accuracy can be established in a wide spectrum of patients.
- Use of individual patient data meta-analysis in diagnostic test accuracy reviews should be explored to allow heterogeneity to be considered in more detail.

Recommendations for those producing and using health technology assessments

The following points are recommended to those producing and using health technology assessments.

- Reviewers should be encouraged to follow recent guidelines on systematic reviews and to use the recently developed Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool to assess.
- Diagnostic test accuracy meta-analyses should not be carried out without the involvement of a statistician familiar with the field.

- Investigators carrying out primary studies should be encouraged to follow the STARD statement to promote the completeness and quality of reporting of test accuracy studies. This should include a clear description of the setting in which the test is being applied.

Acknowledgements

The authors wish to thank their expert advisers for their advice and comments on the draft protocol and draft report. They would also like to thank Penny Whiting of the Centre for Reviews and Dissemination at the University of York for providing the list of eligible reviews from the DARE database.

This report was commissioned by the NHS R&D HTA Programme. This report remains the responsibility of the Southampton Health Technology Assessments Centre, Wessex Institute for Health Research and Development, University of Southampton, and the views expressed are those of the authors and not necessarily those of the NHS R&D Programme. Any errors are the responsibility of the authors.

Contributions of authors

Jacqueline Dinnes (Research Scientist in Evidence Synthesis), Jonathan Deeks (Senior Medical Statistician), Jo Kirby (Research Fellow) and Paul Roderick (Senior Lecturer in Public Health Medicine) all contributed to the conception and design, analysis and interpretation of data, drafting and revision of the report and the final approval of manuscript.

References

1. Alderson P, Green S, Higgins JPT, editors. *Cochrane reviewers' handbook* 4.2.1 [updated December 2003]. *The Cochrane Library*, Issue 1. Chichester: John Wiley; 2004.
2. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *BMJ* 2004;**328**:1040.
3. Whiting P, Rutjes A, Dinnes J, Reitsma J, Bossuyt P, Kleijnen J. Development and validation of methods for assessing the quality and reporting of diagnostic studies. *Health Technol Assess* 2004;**8**(25).
4. Deville WL, Buntinx F. Guidelines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests. In Knottnerus JA, editor. *The evidence base of clinical diagnosis*. London: BMJ Books; 2002. pp. 145–66.
5. Glasziou P, Irwig L, Bain C, Colditz G. *Diagnostic tests. Systematic reviews in health care: a practical guide*. Cambridge: Cambridge University Press; 2001.
6. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, *et al*. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;**120**:667–76.
7. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;**48**:119–30.
8. van der Schouw YT, Verbeek AL, Ruijs SH. Guidelines for the assessment of new diagnostic tests. *Invest Radiol* 1995;**30**:334–40.
9. Deeks JJ. 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*. London: BMJ Books; 2001. pp. 248–82.
10. Shapiro DE. Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. *Acad Radiol* 1995; **2** (Suppl 1):S37–47.
11. Simel D, Samsa G, Matchar D. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol* 1991; **44**:763–70.
12. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002; **21**:1525–37.
13. Deeks J. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001; **323**:157–62.
14. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002; **21**:1539–58.
15. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6.
16. Rutjes A, Reitsma J, Di Nisio M, Smidt N, Zwinderman AH, Rijn JC, *et al*. Bias in diagnostic accuracy studies due to shortcomings in design and conduct. Presented at the XI Annual Cochrane Colloquium, Barcelona, 26–31 October 2003.
17. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987;**6**:411–23.
18. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. *Radiology* 1988;**167**:565–9.
19. Diamond GA. Reverend Bayes' silent majority. An alternative factor affecting sensitivity and specificity of exercise electrocardiography. *Am J Cardiol* 1986;**57**:1175–80.
20. Heffner JE. Evaluating diagnostic tests in the pleural space. Differentiating transudates from exudates as a model. *Clin Chest Med* 1998; **19**:277–93.
21. Mol BW, Bossuyt PMM. Evaluating the effectiveness of diagnostic tests. Tubal subfertility and ectopic pregnancy: evaluating the effectiveness of diagnostic tests. (PhD thesis, Department of Clinical Epidemiology. University of Amsterdam.) 1999.
22. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. *J Clin Epidemiol* 1999;**52**:943–51.
23. Fletcher RH, Fletcher SW, Wagner EH. Studying cases. *Clinical epidemiology: the essentials*. London: Williams & Wilkins; 1996. pp. 208–27.
24. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926–30.
25. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol* 1992;**45**:1143–54.
26. Hlatky MA, Pryor DB, Harrell-FE J, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. *Am J Med* 1984;**77**:64–71.

27. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med* 1997;**16**:981–91.

28. Deeks JJ, Higgins JP, Altman DG. Analysing and presenting results. In Alderson P, Green S, Higgins JPT, editors. *Cochrane reviewers' handbook 4.2.1* [updated December 2003]. *The Cochrane Library*. Issue 1. Chichester: John Wiley; 2004. pp. 68–139.

29. Macaskill, P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol* 2004;**57**(9):925–32.

30. Deeks JJ, Morris JM. Evaluating diagnostic tests. *Ballière's Clin Obstet Gynaecol* 1996;**10**:613–30.

31. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993; **13**:313–21.

32. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data analytic approaches and some additional considerations. *Stat Med* 1993;**12**:1293–316.

33. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988;**7**:889–94.

34. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In Egger M, Davey Smith G, Altman D, editors. *Systematic reviews in health care: meta analysis in context*. London: BMJ Books; 2001. pp. 285–312.

35. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7**:177–88.

36. McCullagh P, Nelder JA. *Generalised linear models*. London: Chapman and Hall; 1989.

37. Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic-test performances – receiver-operating-characteristic summary point estimates. *Medical Decis Making* 1993;**13**:253–7.

38. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**:557–60.

39. Jaeschke R, Guyatt G, Sackett DL. Users guides to the medical literature. VI. How to use an article about a diagnostic test. B: What are the results and will they help me in caring for my patients. *JAMA* 1994;**271**:703–7.

40. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull* 1995; **117**:167–78.

41. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.

42. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 2002;**21**:1237–56.

43. Deville W, Yzermans N, Bouter LM, Bezemer, PD, van der Windt DA. Heterogeneity in systematic reviews of diagnostic studies. In *2nd symposium on systematic reviews: beyond the basics*. Oxford. January. 1999.

44. 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. *Hum Reprod* 1998;**13**:3220–7.

45. Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. *Acad Radiol* 1995;**2** (Suppl 7):S48–56.

46. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;**20**:2865–84.

47. Glas AS. PhD thesis. University of Amsterdam; 2003.

48. van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Stat Med* 1993;**12**:2273–84.

49. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* 2002; **21**:589–624.

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

51. 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.

52. D'Arcy C, McGee S. Does this patient have carpal tunnel syndrome? *JAMA* 2000;**283**:3110–17.

53. 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.

54. Nelson H, Helfand M. Screening for chlamydial infection. *Am J Prev Med* 2001;**20** (3 Suppl):95–107.

55. 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.

56. Adams E, Flynn K. *Positron emission tomography: descriptive analysis of experience with PET in VA*. Technology Assessment Program No. 55. Boston, MA: Health Services Research and Development Services; 1998.

57. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al*. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;**138**:W1–W12.

58. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. *J R Stat Soc A* 1988; **151**:419–63.

59. Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *JAMA* 1992;**263**:374–8.

60. Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt AW, *et al*. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol* 2002;**2**:1–13.

61. Haynes RB, Sackett DL. Purpose and procedure (abbreviated). *Evid Based Med* 1995;**1**:2.

62. Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? *JAMA* 1995;**273**:313–19.

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

64. 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 Commun Health* 1998;**52**:45–52.

65. Sackett DL, Haynes RB, Guyatt GH, Tugwell T. *Clinical epidemiology. A basic science for clinical medicine*. London: Little, Brown; 1991.

66. Bader J, Shugars D, Rozier G, Lohr K, Bonito A, Nelson J. *Diagnosis and management of dental caries*. Evidence Report/Technology Assessment No. 36. Rockville, MD: Agency for Healthcare Research and Quality; 2001.

67. 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.

68. Bafounta M, Beauchet A, Aegerter P, Saiag P. Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma? Results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests. *Arch Dermatol* 2001; **137**:1343–50.

69. Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests. *Recommended methods* [updated 6 June 1996]. URL://som.flinders.edu.au/cochrane/

70. Balk E, Ioannidis J, Salem D, Chew P, Lau J. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: a meta-analysis. *Ann Emerg Med* 2001;**37**:478–94.

71. Lau J, Ioannidis J, Balk E, Milch C, Terrin N, Chew P, *et al*. Diagnosing acute cardiac ischemia in the emergency department: a systematic review of the accuracy and clinical effect of current technologies. *Ann Emerg Med* 2001;**37**:453–60.

72. Banks E. Hormone replacement therapy and the sensitivity and specificity of breast cancer screening: a review. *J Med Screen* 2001;**8**:29–35.

73. Barton M, Harris RD, Fletcher S. Does this patient have breast cancer? *JAMA* 1999;**282**:1270–80.

74. 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.

75. 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.

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

77. Becker D, Philbrick J, Abbitt P. Real-time ultrasonography for the diagnosis of lower extremity deepvenous thrombosis: the wave of the future? *Arch Intern Med* 1989;**149**:1731–4.

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

79. Berger M, Velden JJIM, Lijmer J, de K, Prins A, Bohnen A. Abdominal symptoms: do they predict gallstones? A systematic review. *Scand J Gastroenterol* 2000;**35**:70–6.

80. Berry E, Kelly S, Hutton J, Harris K, Roderick P, Boyce J, *et al*. A systematic literature review of spiral and electron beam computed tomography: with particular reference to clinical applications in hepatic lesions, pulmonary embolus and coronary artery disease. *Health Technol Assess* 1999; **3**(18).

81. 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* 2004;**70**:1028–35.

82. Berry E, Kelly S, Westwood ME, Davies LM, Gough MJ, Bamford JM, *et al*. The cost-effectiveness of magnetic resonance angiography for carotid artery stenosis and peripheral vascular disease: a systematic review. *Health Technol Assess* 2002;**6**(7).

83. Westwood ME, Kelly S, Berry E, Bamford JM, Gough MJ, Airey CM, *et al*. Use of magnetic resonance angiography to select candidates with recently symptomatic carotid stenosis for surgery: systematic review. *BMJ* 2002;**324**:198.

84. Bjelland I, Dahl A, Haug T, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale – an updated literature review. *J Psychosom Res* 2002;**52**:69–77.

85. Blakeley DD, Oddone EZ, Hasselblad V, Simel DL, Matchar DB. Noninvasive carotid artery testing. A meta-analytic review. *Ann Intern Med* 1995; **122**:360–7.

86. 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.

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

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

89. Buchanan A, Leese M. Detention of people with dangerous severe personality disorders: a systematic review. *Lancet* 2001;**358**:1955–9.

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

91. Cabana MD, Alavi A, Berlin JA, Shea JA, Kim CK, Williams SV. Morphine-augmented hepatobiliary scintigraphy: a meta-analysis. *Nucl Med Commun* 1995;**16**:1068–71.

92. Campens D, Buntinx F. Selecting the best renal function tests. A meta-analysis of diagnostic studies. *Int J Technol Assess Health Care* 1997; **13**:343–56.

93. Carlson KJ, Skates SJ, Singer DE. Screening for ovarian cancer. *Ann Intern Med* 1994;**121**:124–32.

94. Cher D, Conwell J, Mandell J. MRI for detecting silicone breast implant rupture: meta-analysis and implications. *Ann Plast Surg* 2001;**47**:367–80.

95. Chesson AL, Ferber RA, Fry JM, Grigg-Damberger M, Hartse KM, Hurwitz TD, *et al*. The indications for polysomnography and related procedures. *Sleep* 1997;**20**:423–87.

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

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

98. Guyatt GH. Critical evaluation of radiologic technologies. *J Can Assoc Radiol* 1992;**43**:6–7.

99. Choi H, Liu S, Merkel P, Colditz G, Niles J. Diagnostic performance of antineutrophil cytoplasmic antibody tests for idiopathic vasculitides: metaanalysis with a focus on antimyoperoxidase antibodies. *J Rheumatol* 2001;**28**:1584–90.

100. Clarke C, Davies P. Systematic review of acute levodopa and apomorphine challenge tests in the diagnosis of idiopathic Parkinson's disease. *J Neurol Neurosurg Psychiatry* 2000;**69**:590–4.

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

102. Cuzick J, Sasieni P, Davies P, Adams J, Normand C, Frater A, *et al*. A systematic review of the role of human papillomavirus testing within a cervical screening programme. *Health Technol Assess* 1999; **3**(14).

103. 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.

104. 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.

105. de Bruyn G, Graviss E. A systematic review of the diagnostic accuracy of physical examination for the detection of cirrhosis. *BMC Med Inform Decis Mak* 2001;**1**:6.

106. 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.

107. Deville WL, van der Windt DA, Dzaferagic A, Bezemer PD, Bouter LM. The test of Lasegue: systematic review of the accuracy in diagnosing herniated discs. *Spine* 2000;**25**:1140–7.

108. 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.

109. Dharnidharka VR, Kwon C, Stevens G. Serum cystatin C is superior to serum creatinine as a marker of kidney function: a meta-analysis. *Am J Kidney Dis* 2002;**40**:221–6.

110. Di Fabio RP. Meta-analysis of the sensitivity and specificity of platform posturography. *Arch Otolaryngol Head Neck Surg* 1996;**122**:150–6.

111. Dinnes J, Moss S, Melia J, Blanks R, Song F, Kleijnen J. Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. *Breast* 2001;**10**:455–63.

112. Divakaran T, Waugh J, Clark T, Khan K, Whittle M, Kilby M. Noninvasive techniques to detect fetal anemia due to red blood cell alloimmunization: a systematic review. *Obstet Gynecol* 2001;**98**:509–17.

113. Ebell MH, Flewelling D, Flynn CA. A systematic review of troponin T and I for diagnosing acute myocardial infarction. *J Fam Pract* 2000;**49**:550–6.

114. Eberhard-Gran M, Eskild A, Tambs K, Opjordsmoen S, Samuelsen S. Review of validation studies of the Edinburgh postnatal depression scale. *Acta Psychiatr Scand* 2001;**104**:243–9.

115. ECRI (Emergency Care Research Institute). *Diagnosis and treatment of swallowing disorders (dysphagia) in acute-care stroke patients*. Evidence Report/Technology Assessment No. 8. Rockville, MD: Agency for Health Care Policy and Research; 1999.

116. Eden K, Mahon S, Helfand M. Screening high-risk populations for thyroid cancer. *Med Pediatr Oncol* 2001;**36**:583–91.

117. Eiberg J, Lundorf E, Thomsen C, Schroeder T. Peripheral vascular surgery and magnetic resonance arteriography – a review. *Eur J Vasc Endovasc Surg* 2001;**22**:396–402.

118. Ernst E. Iridology: a systematic review. *Forsch Komplementarmed* 1999;**6**:7–9.

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

120. Faron G, Boulvain M, Irion O, Barnard PM, Fraser WD. Prediction of preterm delivery by fetal fibronectin: a meta-analysis. *Obstet Gynecol* 1998; **92**:153–8.

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

122. Fiorino AS. Electron-beam computed tomography, coronary artery calcium, and evaluation of patients with coronary artery disease. *Ann Intern Med* 1998; **128**:839–47.

123. Fischer B, Mortensen J, Hojgaard L. Positron emission tomography in the diagnosis and staging of lung cancer: a systematic, quantitative review. *Lancet Oncol* 2001;**2**:659–66.

124. Fleischmann KE, Hunink MG, Kuntz KM, Douglas PS. Exercise echocardiography of exercise SPECT imaging: a meta analysis of diagnostic test performance. *JAMA* 1998;**280**:913–20.

125. Fowlie PW, Schmidt B. Diagnostic tests for bacterial infection from birth to 90 days – a systematic review. *Arch Dis Child* 1998; **78**:F92–F98.

126. Frost SC, Amendola A. Is stress radiography necessary in the diagnosis of acute or chronic ankle instability. *Clin J Sport Med* 1999;**9**:40–5.

127. Garzon P, Eisenberg MJ. Functional testing for the detection of restenosis after percutaneous transluminal coronary angioplasty: a meta-analysis. *Can J Cardiol* 2001;**17**:41–8.

128. Gianrossi R, Detrano R, Colombo A, Froelicher V. Cardiac fluoroscopy for the diagnosis of coronary artery disease: a meta analytic review. *Am Heart J* 1990;**120**:1179–88.

129. Wachter KW. Disturbed by meta-analysis? *Science* 1988;**241**:1407–8.

130. Gifford DR, Holloway RG, Vickrey BG. Systematic review of clinical prediction rules for neuroimaging in the evaluation of dementia. *Arch Intern Med* 2000;**160**:2855–62.

131. Gottlieb RH, Widjaja J, Tian L, Rubens DJ, Voci SL. Calf sonography for detecting deep venous thrombosis in symptomatic patients: experience and review of the literature. *J Clin Ultrasound* 1999;**27**:415–20.

132. Gould M, Maclean C, Kuschner W, Rydzak C, Owens D. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. *JAMA* 2001; **285**:914–24.

133. 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.

134. Gronseth GS, Ashman EJ. Practice parameter: the usefulness of evoked potentials in identifying clinically silent lesions in patients with suspected multiple sclerosis (an evidence-based review): Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology* 2000;**54**:1720–5.

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

136. Harvey SA, Black KJ. The dexamethasone suppression test for diagnosing depression in stroke patients. *Ann Clin Psychiatry* 1996;**8**:35–9.

137. 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.

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

139. Helfand M, Mahon S, Eden K, Frame PS, Orleans TC. Screening for skin cancer. *Am J Prev Med* 2001;**20**:47–58.

140. Hider P, Nicholas B. The early detection and diagnosis of breast cancer: an update. *N Z Health Technol Assessment Rep* 1999;**2**:1–150.

141. New Zealand Health Technology Assessment Centre. Diagnostic tests. In: NHMRC, editor. *How to review the evidence: systematic identification and review of the scientific literature; handbook series on preparing clinical practice guidelines*. Canberra: Biotext; 1999. pp. 75–88.

142. 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).

143. 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.

144. Hobby J, Tom B, Bearcroft P, Dixon A. Magnetic resonance imaging of the wrist: diagnostic performance statistics. *Clin Radiol* 2001;**56**:50–7.

145. Hoffman RM, Clanon DL, Littenberg B, Frank JJ, Peirce JC. Using the free-to-total prostate-specific antigen ratio to detect prostate cancer in men with nonspecific elevations of prostate-specific antigen levels. *J Gen Intern Med* 2000;**15**:739–48.

146. Hoffman RM, Clanon DL, Chavez M, Peirce JC. Using multiple cutpoints for the free-to-total prostate specific antigen ratio improves the accuracy of prostate cancer detection. *Prostate* 2002;**52**:150–8.

147. Hofman PA, 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.

148. Hooft L, Hoekstra O, Deville W, Lips P, Teule G, Boers M, *et al*. Diagnostic accuracy of 18F-fluorodeoxyglucose positron emission tomography in the follow-up of papillary or follicular thyroid cancer. *J Clin Endocrinol Metab* 2001;**86**:3779–86.

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

150. Huicho L, Campos-Sanchez M, Alamo C. Metaanalysis of urine screening tests for determining the risk of urinary tract infection in children. *Pediatr Infect Dis J* 2002;**21**:1–11.

151. 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.

152. Hurley JC. Concordance of endotoxemia with Gram-negative bacteremia. A meta-analysis using receiver operating characteristic curves. *Arch Pathol Lab Med* 2000;**124**:1157–64.

153. Ioannidis J, Lau J. Technical report: Evidence for the diagnosis and treatment of acute uncomplicated sinusitis in children: a systematic overview. *Pediatrics* 2001;**108**:E57.

154. Ioannidis J, Salem D, Chew P, Lau J. Accuracy and clinical effect of out-of-hospital electrocardiography in the diagnosis of acute cardiac ischemia: a meta-analysis. *Ann Emerg Med* 2001;**37**:461–70.

155. Ioannidis J, Salem D, Chew P, Lau J. Accuracy of imaging technologies in the diagnosis of acute cardiac ischemia in the emergency department: a meta-analysis. *Ann Emerg Med* 2001;**37**:471–7.

156. Kallmes DF, Omary RA, Dix JE, Evans AJ, Hillman BJ. Specificity of MR angiography as a confirmatory test of carotid artery stenosis. *Am J Neuroradiol* 1996;**17**:1501–6.

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

158. Kim C, Kwok Y, Heagerty P, Redberg R. Pharmacologic stress testing for coronary disease diagnosis: a meta-analysis. *Am Heart J* 2001; **142**:934–44.

159. Kinkel K, Hricak H, Lu Y, Tsuda K, Filly RA. US characterization of ovarian masses: a meta-analysis. *Radiology* 2000;**217**:803–11.

160. Kinkel K, Kaji Y, Yu KK, Segal MR, Lu Y, Powell CB, *et al*. Radiologic staging in patients with endometrial cancer: a meta-analysis. *Radiology* 1999;**212**:711–18.

161. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. *Lancet Oncol* 2002;**3**:159–65.

162. Klompas M. Does this patient have an acute thoracic aortic dissection? *JAMA* 2002; **287**:2262–72.

163. Knopman DS, DeKosky S, Cummings J, Chui H, Corey-Bloom J, Relkin N, *et al*. Practice parameter: diagnosis of dementia (an evidence-based review) – Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology* 2001;**56**:1143–53.

164. 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.

165. Koelemay M, Lijmer J, Stoker J, Legemate D, Bossuyt P. Magnetic resonance angiography for the evaluation of lower extremity arterial disease: a meta-analysis. *JAMA* 2001;**285**:1338–45.

166. Kowalski J, Tu XM, Jia G, Pagano M. A comparative meta-analysis on the variability in test performance among FDA-licensed enzyme immunosorbent assays for HIV antibody testing. *J Clin Epidemiol* 2001;**54**:448–61.

167. 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.

168. Kwok Y, Kim C, Grady D, Segal M, Redberg R. Meta-analysis of exercise testing to detect coronary artery disease in women. *Am J Cardiol* 1999; **83**:660–6.

169. 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.

170. Lau J, Zucker D, Engels E, Balk E, Barza M, Terrin N, *et al*. *Diagnosis and treatment of acute bacterial rhinosinusitis*. Evidence Report/Technology Assessment No. 9. Rockville, MD: Agency for Health Care Policy and Research; 1999.

171. Engels EA, Terrin N, Barza M, Lau J. Meta-analysis of diagnostic tests for acute sinusitis. *J Clin Epidemiol* 2000;**53**:852–62.

172. Law J, Boyle J, Harris F, Harkness A, Nye C. Screening for speech and language delay: a systematic review of the literature. *Health Technol Assess* 1998;**2**(9).

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

174. Leitich H, Egarter C, Kaider A, Hohlagschwandtner M, Berghammer P, Husslein P. Cervicovaginal fetal fibronectin as a marker for preterm delivery: a meta-analysis. *Am J Obstet Gynecol* 1999;**180**:1169–76.

175. Li J. Capnography alone is imperfect for endotracheal tube placement confirmation during emergency intubation. *J Emerg Med* 2001;**20**:223–9.

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

177. Lindbaek M, Hjortdahl P. The clinical diagnosis of acute purulent sinusitis in general practice – a review. *Br J Gen Pract* 2002;**52**:491–5.

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

179. 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.

180. Jaeschke R, Guyatt G, Sackett DL. Users guides to the medical literature VI. How to use an article about a diagnostic test. A: are the results of the study valid? *JAMA* 1994;**271**:389–91.

181. Lysakowski C, Walder B, Costanza M, Tramer M. Transcranial Doppler versus angiography in patients with vasospasm due to a ruptured cerebral aneurysm – a systematic review. *Stroke* 2001; **32**:2292–8.

182. Mackenzie R, Palmer CR, Lomas DJ, Dixon AK. Magnetic resonance imaging of the knee: diagnostic performance statistics. *Clin Radiol* 1996; **51**:251–7.

183. Mango LJ, Radensky PW. Interactive neural-network-assisted screening: a clinical assessment. *Acta Cytol* 1998;**42**:233–45.

184. Markert RJ, Walley ME, Guttman TG, Mehta R. A pooled analysis of the Ottawa ankle rules used on adults in the ED. *Am J Emerg Med* 1998; **16**:564–7.

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

186. McCrory D, Matchar D, Bastian L, Datta S, Hasselblad V, Hickey J, *et al*. *Evaluation of cervical cytology*. Evidence Report/Technology Assessment No. 36. Rockville, MD: Agency for Healthcare Research and Quality; 1999.

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

188. McNaughton-Collins M, MacDonald R, Wilt TJ. Diagnosis and treatment of chronic abacterial prostatitis: a systematic review. *Ann Intern Med* 2000;**133**:367–81.

189. Merritt RM, Williams MF, James TH, Porubsky ES. Detection of cervical metastasis. A meta-analysis comparing computed tomography with physical examination. *Arch Otolaryngol Head Neck Surg* 1997;**123**:149–52.

190. 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.

191. Mitchell MF, Cantor SB, Ramanujam N, Tortolero-Luna G, Richards-Kortum R. Fluorescence spectroscopy for diagnosis of squamous intraepithelial lesions of the cervix. *Obstet Gynecol* 1999;**93**:462–70.

192. Mitchell MF, Schottenfeld D, Tortolero-Luna G, Cantor SB, Richards-Kortum R. Colposcopy for the diagnosis of squamous intraepithelial lesions: a meta-analysis. *Obstet Gynecol* 1998;**91**:626–31.

193. 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.

194. 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. *Fertil Steril* 1997; **67**:1031–7.

195. 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.

196. Mol BW, Meijer S, Yuppa S, Tan E, de Vries J, Bossuyt PM, *et al*. Sperm penetration assay in predicting successful *in vitro* fertilization: a meta-analysis. *J Reprod Med* 1998;**43**:503–8.

197. Australasian Cochrane Centre. *Oto-acoustic emission audiometry*. Canberra, ACT: Medicare Services Advisory Committee; 1999.

198. 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.

199. Muris JWM, Starmans R, Pop P, Crebolder HFJM, Knottnerus JA. Discriminant value of symptoms in patients with dyspepsia. *J Fam Pract* 1994;**38**:139–43.

200. Muris JW, Starmans R, Pop P, Crebolder HF, Knottnerus JA. The diagnostic value of symptoms for the identification of patients with an increased risk of colorectal disease. A criteria-based analysis. *Fam Pract* 1992;**9**:415–20.

201. Mushlin AI, Kouides RW, Shapiro DE. Estimating the accuracy of screening mammography: a meta-analysis. *Am J Prevent Med* 1998;**14**:143–53.

202. Mustafa BO, Rathbun SW, Whitsett TL, Raskob GE. Sensitivity and specificity of ultrasonography in the diagnosis of upper extremity deep vein thrombosis: a systematic review. *Arch Intern Med* 2002;**162**:401–4.

203. Nallamothu B, Saint S, Bielak L, Sonnad S, Peyser P, Rubenfire M, *et al*. Electron-beam computed tomography in the diagnosis of coronary artery disease. *Arch Intern Med* 2001; **161**:833–8.

204. Nanda K, McCrory DC, Myers ER, Bastian LA, Hasselblad V, Hickey JD, *et al*. Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: a systematic review. *Ann Intern Med* 2000;**132**:810–19.

205. 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.

206. Oei SG, Helmerhorst FM, Keirse MJ. When is the post-coital test normal? A critical appraisal. *Hum Reprod* 1995;**10**:1711–14.

207. Olatidoye AG, Wu AH, Feng YJ, Waters D. Prognostic role of troponin T versus troponin I in unstable angina pectoris for cardiac events with meta-analysis comparing published studies. *Am J Cardiol* 1998;**81**:1405–10.

208. Oosterhuis WP, Niessen RW, Bossuyt PM, Sanders GT, Sturk A. Diagnostic value of the mean corpuscular volume in the detection of vitamin B12 deficiency. *Scand J Clin Lab Invest* 2000; **60**:9–18.

209. Orr RK, Porter D, Hartman D. Ultrasonography to evaluate adults for appendicitis: decision making based on meta-analysis and probabilistic reasoning. *Acad Emerg Med* 1995;**2**:644–50.

210. 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.

211. Pasternack I, Tuovinen E, Lohman M, Vehmas T, Malmivaara A. MR findings in humeral epicondylitis: a systematic review. *Acta Radiol* 2001; **42**:434–40.

212. Patel SR, Wiese W, Patel SC, Ohl C, Byrd JC, Estrada CA. Systematic review of diagnostic tests for vaginal trichomoniasis. *Infect Dis Obstet Gynecol* 2000;**8**:248–57.

213. Dijkhuizen FP, Mol BW, Brolmann HA, Heintz AP. The accuracy of endometrial sampling in the diagnosis of patients with endometrial carcinoma and hyperplasia: a meta-analysis. *Cancer* 2000; **89**:1765–72.

214. 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.

215. Peters AL, Davidson MB, Schriger DL, Hasselblad V. A clinical approach for the diagnosis of diabetes mellitus: an analysis using glycosylated hemoglobin levels. Meta-analysis Research Group on the Diagnosis of Diabetes Using Glycated Hemoglobin Levels [published erratum appears in *JAMA* 1997;**277**:1125]. *JAMA* 1996;**276**:1246–52.

216. Petersen R, Stevens J, Ganguli M, Tangalos E, Cummings J, DeKosky S. Practice parameter: early detection of dementia: mild cognitive impairment (an evidence-based review) – Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology* 2001;**56**:1133–42.

217. 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.

218. Rao G. Diagnostic yield of screening for type 2 diabetes in high-risk patients: a systematic review. *J Fam Pract* 1999;**48**:805–10.

219. Rao SC, Fehlings MG. The optimal radiologic method for assessing spinal canal compromise and cord compression in patients with cervical spinal cord injury: part 1: an evidence-based analysis of the published literature. *Spine* 1999;**24**:598–604.

220. Rathbun SW, Raskob GE, Whitsett TL. Sensitivity and specificity of helical computed tomography in the diagnosis of pulmonary embolism: a systematic review. *Ann Intern Med* 2000;**132**:227–32.

221. 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.

222. Revah A, Hannah ME, Sue-A-Quan AK. Fetal fibronectin as a predictor of preterm birth: an overview. *Am J Perinatol* 1998;**15**:613–21.

223. Ross SD, Allen IE, Harrison KJ, Kvasz M, Connelly J, Sheinhait IA. *Systematic review of the literature regarding the diagnosis of sleep apnea*. Evidence Report/Technology Assessment No. 1. Rockville, MD: Agency for Health Care Policy and Research; 1999.

224. Safriel Y, Zinn H. CT pulmonary angiography in the detection of pulmonary emboli: a meta-analysis of sensitivities and specificities. *Clin Imaging* 2002;**26**:101–5.

225. Scheid D, McCarthy L, Lawler F, Hamm R, Reilly K. Screening for microalbuminuria to prevent nephropathy in patients with. *J Fam Pract* 2001;**50**:661–8.

226. McKibbon A, Walker-Dilks CJ. *Evidence-based medicine for librarians: panning for gold. How to apply research methodology to search for therapy, diagnosis, etiology, and prognosis articles*. Presented at the MLA Annual Meeting, Washington, DC, 1995.

227. Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer: a meta-analysis. *JAMA* 1997;**278**:1096–101.

228. Schwimmer J, Essner R, Patel A, Jahan SA, Shepherd JE, Park K, *et al*. A review of the literature for whole-body FDG PET in the management of patients with melanoma. *Q J Nucl Med* 2000;**44**:153–67.

229. Scouller K, Conigrave KM, Macaskill P, Irwig L, Whitfield JB. Should we use carbohydrate-deficient transferrin instead of gamma-glutamyltransferase for detecting problem drinkers? A systematic review and metaanalysis. *Clin Chem* 2000;**46**:1894–902.

230. Selley S, Donovan J, Faulkner A, Coast J, Gillatt D. Diagnosis, management and screening of early localised prostate cancer. *Health Technol Assess* 1997;**1**(2).

231. Siegman-Igra Y, Anglim AM, Shapiro DE, Adal KA, Strain BA, Farr BM. Diagnosis of vascular catheter-related bloodstream infection: a meta-analysis. *J Clin Microbiol* 1997;**35**:928–36.

232. Smith ER, Peterson J, Okorodudu AO, Bissell MG. Does the addition of unconjugated estriol in maternal serum screening improve the detection of trisomy 21? *Clin Lab Manage Rev* 1996; Mar/Apr: 176–81.

233. Smith-Bindman R, Hosmer W, Feldstein V, Deeks J, Goldberg J. Second-trimester ultrasound to detect fetuses with down syndrome: a meta-analysis. *JAMA* 2001;**285**:1044–55.

234. Smith-Bindman R, Kerlikowske K, Feldstein VA, Subak L, Scheidler J, Segal M, *et al*. Endovaginal ultrasound to exclude endometrial cancer and other endometrial abnormalities. *JAMA* 1998; **280**:1510–17.

235. Solomon DH, Simel DL, Bates DW, Katz JN, Schaffer JL. The rational clinical examination. Does this patient have a torn meniscus or ligament of the knee? Value of the physical examination. *JAMA* 2001;**286**:1610–20.

236. Sonnad SS, Langlotz CP, Schwartz JS. Accuracy of MR imaging for staging prostate cancer: a meta-analysis to examine the effect of technologic change. *Acad Radiol* 2001;**8**:149–57.

237. Spencer-Green G, Alter D, Welch HG. Test performance in systemic sclerosis: anti-centromere and anti-Scl-70 antibodies. *Am J Med* 1997; **103**:242–8.

238. Stengel D, Bauwens K, Sehouli J, Porzsolt F, Rademacher G, Mutze S, *et al*. Systematic review and meta-analysis of emergency ultrasonography for blunt abdominal trauma. *Br J Surg* 2001; **88**:901–12.

239. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine: how to practice and teach EBM*. Churchill Livingstone: Edinburgh; 2000.

240. Storgaard H, Nielsen SD, Gluud C. The validity of the Michigan Alcoholism Screening Test (MAST). *Alcohol Alcohol* 1994;**29**:493–502.

241. 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.

242. Taylor-Weetman K, Wake B, Hyde C. *Comparison of panoramic and bitewing radiography for the detection of dental caries: a systematic review of diagnostic tests*. Birmingham: University of Birmingham, Department of Public Health and Epidemiology; 2002.

243. Tenner S, Dubner H, Steinberg W. Predicting gallstone pancreatitis with laboratory parameters: a meta-analysis. *Am J Gastroenterol* 1994;**89**:1863–6.

244. Helfand M, Thompson D, Davis R, McPhillips H, Homer C, Lieu T. *Newborn hearing screening*. Systematic Evidence Review No. 5. Rockville, MD: Agency for Healthcare Research and Quality; 2001.

245. US Preventive Services Task Force. *Guide to clinical preventive services: an assessment of the effectiveness of 169 interventions*. Baltimore, MD: Williams & Wilkins; 1989.

246. 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.

247. van Beek EJ, Brouwers EM, Song B, Bongaerts AH, Oudkerk M. Lung scintigraphy and helical computed tomography for the diagnosis of pulmonary embolism: a meta-analysis. *Clin Appl Thromb Hemost* 2001;**7**:87–92.

248. 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.

249. van der Wurff P, Meyne W, Hagmeijer RH. Clinical tests of the sacroiliac joint. *Man Ther* 2000;**5**:89–96.

250. van der Wurff P, Hagmeijer RH, Meyne W. Clinical tests of the sacroiliac joint. A systemic methodological review. Part 1: reliability. *Man Ther* 2000;**5**:30–6.

251. Varonen H, Makela M, Savolainen S, Laara E, Hilden J. Comparison of ultrasound, radiography, and clinical examination in the diagnosis of acute maxillary sinusitis: a systematic review. *J Clin Epidemiol* 2000;**53**:940–8.

252. Vasbinder GB, Nelemans PJ, Kessels AG, Kroon AA, de Leeuw PW, van Engelshoven JM. Diagnostic tests for renal artery stenosis in patients suspected of having renovascular hypertension: a meta-analysis. *Ann Intern Med* 2001;**135**:401–11.

253. Verkooijen HM, Peeters PH, Buskens E, Koot VC, Borel-Rinkes IH, Mali WP, *et al*. Diagnostic accuracy of large-core needle biopsy for nonpalpable breast disease: a meta-analysis. *Br J Cancer* 2000;**82**:1017–21.

254. Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US – a meta-analysis. *Radiology* 2000;**216**:67–77.

255. Vroomen PC, de Krom MC, Knottnerus JA. Diagnostic value of history and physical examination in patients suspected of sciatica due to disc herniation: a systematic review. *J Neurol* 1999;**246**:899–906.

256. Watson EJ, Templeton A, Russell I, Paavonen J, Mardh PA, Stary A, *et al*. The accuracy and efficacy of screening tests for *Chlamydia trachomatis*: a systematic review. *J Med Microbiol* 2002; **51**:1021–31.

257. 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.

258. White PM, Wardlaw JM, Easton V. Can noninvasive imaging accurately depict intracranial aneurysms? A systematic review. *Radiology* 2000;**217**:361–70.

259. Whited JD, Grichnik JM. Does this patient have a mole or a melanoma? *JAMA* 1998;**279**: 696–701.

260. Whitsel EA, Boyko EJ, Siscovick DS. Reassessing the role of QTc in the diagnosis of autonomic failure among patients with diabetes: a meta-analysis. *Diabetes Care* 2000;**23**:241–7.

261. Wiese W, Patel SR, Patel SC, Ohl CA, Estrada CA. A meta-analysis of the Papanicolaou smear and wet mount for the diagnosis of vaginal trichomoniasis. *Am J Med* 2000;**108**:301–8.

262. Wijnberger LD, Huisjes AJ, Voorbij HA, Franx A, Bruinse HW, Mol BW. The accuracy of lamellar body count and lecithin/sphingomyelin ratio in the prediction of neonatal respiratory distress syndrome: a meta-analysis. *Br J Obstet Gynaecol* 2001;**108**:583–8.

263. Williams J-WJ, Noel PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? *JAMA* 2002;**287**:1160–70.

264. Davis A, Bamford J, Wilson I, Ramkalawan T, Forshaw M, Wright S. A critical review of the role of neonatal hearing screening in the detection of congenital hearing impairment. *Health Technol Assess* 1997;**1**(10).

Appendix 1

Calculation of diagnostic accuracy statistics

Participants

Test results	Diseased	Non-diseased	
+ ve	True positives *a*	*b* False positives	Total positive
– ve	False negatives *c*	*d* True negatives	Total negative
	Total diseased	Total non-diseased	

Sensitivity	Proportion of diseased who have positive test results	True positives/total diseased $a/(a + c)$
Specificity	Proportion of non-diseased who have negative test results	True negatives/total non-diseased $d/(b + d)$
Positive predictive value (PPV)	Proportion with positive test result who actually have the disease	True positives/total positive $a/(a + b)$
Negative predictive value (NPV)	Proportion with negative test result who really do not have the disease	True negatives/total negative $d/(c + d)$
Positive likelihood ratio (LR +ve)	Likelihood of a person with disease having a positive test result than a person without disease	(True positives/total diseased)/ (false positives/total non-diseased) sensitivity/(1 – specificity)
Negative likelihood ratio (LR –ve)	Likelihood of a person with disease having a negative test result than a person without disease	(False positives/total diseased)/ (true negatives/total non-diseased) (1 – sensitivity)/specificity
Diagnostic odds ratio (DOR)	Odds of positivity among diseased persons, divided by the odds of positivity among non-diseased persons	Positive likelihood ratio/negative likelihood ratio $(a \times d)/(c \times b)$

Appendix 2

Data extraction form used

Systematic Reviews of Diagnostic Tests

☐ Data extracted ☐ Format revised

Study ID: 1 ☐ Select JacqProj ☐ Double checked

Author: Year:

Diagnostic test evaluated:

Gold Standard:

Disease tested for:

Total Studies: Total Patients:

Range Study Size:

Study Comment

SEARCH STRATEGY USED

☐ MEDLINE ☐ Search Hand ☐ Grey literature Language:

☐ EMBASE ☐ Reference lists ☐ Other sources Duration:

☐ Other dbase ☐ Contact authors/expects

INCLUSION CRITERIA

☐ Population

☐ Tests used

☐ Particular outcomes reported

☐ Sample size

☐ Study design

☐ Data for 2 × 2 table

INCLUSION CRITERIA (METHODOLOGICAL)

☐ Appropriate reference test ☐ Complete FU ☐ Incorporation bias

☐ Consecutive or random sample ☐ Blinding used

☐ Prospective studies only ☐ Adequate description of sample

☐ Avoidance of verification bias Other methodological criteria

Comment on inclusion criteria

VALIDITY ASSESSMENT VA tool used:

☐ Adequate description of tests ☐ Avoidance of partial verification bias ☐ Adequate sample description

☐ Approp reference test used ☐ Avoidance of differential verification bias ☐ Approp spectrum included

☐ Method of patient enrolment ☐ Completeness of FU ☐ Treatment paradox

☐ Prospective/Retrospective design ☐ Blinded test interpretation ☐ Uninterpretable results

Other VA criteria:

List items relating to spectrum and/or

VA Comment:

DATA EXTRACTION

Was the use of different thresholds considered in the data extraction?

SYNTHESIS METHODS Synthesis:

☐ Pooled sens/spec ☐ Pooled LRs ☐ Summary ROC curve ☐ Regression ☐ Narrative only

Other methods of pooling:

Comment on synthesis method

How was statistical heterogeneity considered?

0

Stat Test: No Stat Test detail: Stat Graph Detail:

Stat Test Result:

How were threshold effects considered statistically?:

0

☐ Thresh Test: Thresh Test detail:

Thresh Result:

How was heterogeneity investigated:

Which quality or study design effects were considered?

0

Which clinical differences were investigated? (relating to spectrum and setting)

0

What differences in the tests used were investigated?

RESULTS:

Overall comment on review in terms of heterogeneity

Heterogeneity likely in terms of:

List of Poss Influences:

.

Appendix 3

List of excluded reviews

1. Almekinders LC, Temple JD. Etiology, diagnosis, and treatment of tendonitis: an analysis of the literature. *Med Sci Sports Exerc* 1998;**30**:1183–90. Reason for exclusion: not assessing diagnostic accuracy
2. Anderson LA, Janes GR, Jenkins C. Implementing preventive services: to what extent can we change provider performance in ambulatory care: a review of the screening, immunization, and counseling literature. *Ann Behav Med* 1998;**20**:161–7. Reason for exclusion: not assessing diagnostic accuracy
3. Arbyn M, Schenck U. Detection of false negative pap smears by rapid reviewing – a metaanalysis. *Acta Cytol* 2000;**44**:949–57. Reason for exclusion: evaluates a technique for quality control not specifically aimed at diagnosis
4. Austoker J. Screening and self examination for breast cancer. *BMJ* 1994;**309**:168–74. Reason for exclusion: not assessing diagnostic accuracy
5. Bachmann M, Nelson S. *Screening for diabetic retinopathy: a quantitative overview of the evidence, applied to the populations of health authorities and boards*. Bristol: University of Bristol, Department of Social Medicine, Health Care Evaluation Unit; 1996. Reason for exclusion: same as Bachmann, 1998[64]
6. Barlow J, Stewart-Brown S, Fletcher J. Systematic review of the school entry medical examination. *Arch Dis Child* 1998;**78**:301–11. Reason for exclusion: not diagnostic accuracy
7. Bax JJ, Wijns W, Cornel JH, Visser FC, Boersma E, Fioretti PM. Accuracy of currently available techniques for prediction of functional recovery after revascularization in patients with left ventricular dysfunction due to chronic coronary artery disease: comparison of pooled data. *J Am Coll Cardiol* 1997;**30**:1451–60. Reason for exclusion: not assessing diagnostic accuracy
8. Black K, Shea C, Dursun S, Kutcher S. Selective serotonin reuptake inhibitor discontinuation syndrome: proposed diagnostic criteria (Review). *J Psychiatry Neurosci* 2000;**25**:255–61. Reason for exclusion: not diagnostic accuracy
9. Brewer DA, Fung CL, Chapuis PH, Bokey EL. Should relatives of patients with colorectal cancer be screened? A critical review of the literature. *Dis Colon Rectum* 1994;**37**:1328–38. Reason for exclusion: not assessing diagnostic accuracy
10. Brown N. Exploration of diagnostic techniques for malignant melanoma: an integrative review. *Clin Excell Nurse Pract* 2000;**4**:263–71. Reason for exclusion: not systematic review – only details search and inclusion criteria
11. Brumback BA, Holmes LB, Ryan LM. Adverse effects of chorionic villus sampling: a meta-analysis. *Stat Med* 1999;**18**:2163–75. Reason for exclusion: not assessing diagnostic accuracy
12. Buntinx F, Knottnerus J, Andre J, Crebolder HF, Essed GG. The effect of different sampling devices on the presence of endocervical cells in cervical smears: a systematic literature review. *Eur J Cancer Prev* 1994;**3**:23–30. Reason for exclusion: not assessing diagnostic accuracy
13. Buntinx F, Brouwers M. Relation between sampling device and detection of abnormality in cervical smears. A meta-analysis of randomised and quasi-randomised studies. *BMJ* 1996; **313**:1285–90. Reason for exclusion: not assessing diagnostic accuracy
14. Bushnell C, Goldstein L. Diagnostic testing for coagulopathies in patients with ischemic stroke (Review). *Stroke* 2000;**31**:3067–78. Reason for exclusion: not assessing test accuracy *per se*
15. Carter T, Jordan R, Cummins C. *Electrodiagnostic techniques: in the pre-surgical assessment of patients with carpal tunnel syndrome*. Birmingham: Development and Evaluation Service, Department of Public Health and Epidemiology, University of Birmingham; 2000. pp. 1–30. Reason for exclusion: use as prognostic tool; not assessing accuracy
16. Centre for Reviews and Dissemination. Screening for osteoporosis to prevent fractures. *Effect Health Care* 1992;**9**:12. Reason for exclusion: not assessing diagnostic accuracy
17. Centre for Reviews and Dissemination. The management of menorrhagia: what are effective ways of treating excessive regular menstrual blood loss in primary and secondary care? *Effect Health Care* 1995;**9**:1–14. Reason for exclusion: not assessing diagnostic accuracy
18. Centre for Reviews and Dissemination. *Review of the research on the effectiveness of health service interventions to reduce variations in health*. York: University of York, NHS Centre for Reviews and Dissemination; 1995. Reason for exclusion: not assessing diagnostic accuracy
19. Cloft HJ, Joseph GJ, Dion JE. Risk of cerebral angiography in patients with subarachnoid hemorrhage, cerebral aneurysm, and

arteriovenous malformation: a meta-analysis. *Stroke* 1999;**30**:317–20. Reason for exclusion: not assessing diagnostic accuracy

20. Coetzee K, Kruge TF, Lombard CJ. Predictive value of normal sperm morphology: a structured literature review. *Hum Reprod Update* 1998; **4**:73–82. Reason for exclusion: not diagnostic accuracy

21. Cole MG. Impact of geriatric home screening services on mental state: a systematic review. *Int Psychogeriatr* 1998;**10**:97–102. Reason for exclusion: not assessing diagnostic accuracy

22. Coley CM, Barry MJ, Fleming C, Mulley AG. Early detection of prostate cancer part 1: prior probability and effectiveness of tests. *Ann Intern Med* 1997;**126**:394–406. Reason for exclusion: not assessing diagnostic accuracy

23. Conn D, Lieff S. Diagnosing and managing delirium in the elderly. *Can Fam Phys* 2001; **47**:101–8. Reason for exclusion: not a systematic review

24. Davis A, Bamford J, Wilson I, Ramkalawan T, Forshaw M, Wright S. A critical review of the role of neonatal hearing screening in the detection of congenital hearing impairment. *Health Technol Assess* 1997;**1**(10). Reason for exclusion: limited methods presented; not primarily accuracy

25. Davis A, Bamford J, Stevens J. Performance of neonatal and infant hearing screens: sensitivity and specificity. *Br J Audiol* 2001;**35**:3–15. Reason for exclusion: duplicate publication – see Davis, 1997[264]

26. Deeks J. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001; **323**:157–62. Reason for exclusion: not assessing diagnostic accuracy

27. Delaney BC, Hyde C, McManus RJ, Wilson S, Fitzmaurice DA, Jowett S, *et al*. Systematic review of near patient test evaluations in primary care. *BMJ* 1999;**319**:824–7. Reason for exclusion: duplicate of Hobbs, 1997[142]

28. Dumler J. Molecular diagnosis of Lyme disease: review and meta-analysis. *Mol Diagn* 2001;**6**:1–11. Reason for exclusion: not fully systematic review – only reports search

29. Ebrahim S. Detection, adherence and control of hypertension for the prevention of stroke: a systematic review. *Health Technol Assess* 1998;**2**(11). Reason for exclusion: not assessing diagnostic accuracy

30. ECRI. *Analysis of the published literature on device reuse*. Philadelphia, PA: Emergency Care Research Institute; 1997. Reason for exclusion: not assessing diagnostic accuracy

31. Engels EA, Terrin N, Barza M, Lau J. Meta-analysis of diagnostic tests for acute sinusitis. *J Clin Epidemiol* 2000;**53**:852–62. Reason for exclusion: duplicate of Lau, 1999[170]

32. Estrada CA, Bloch RM, Antonacci D, Basnight LL, Patel SR, Patel SC, *et al*. Reporting and concordance of methodologic criteria between abstracts and articles in diagnostic test studies. *J Gen Intern Med* 2000;**15**:183–7. Reason for exclusion: not assessing diagnostic accuracy

33. Federman DG, Concato J, Kirsner RS. Comparison of dermatologic diagnoses by primary care practitioners and dermatologists: a review of the literature. *Arch Fam Med* 1999;**8**:170–2. Reason for exclusion: not diagnostic accuracy – examines % of correct diagnoses only between primary care and dermatology

34. Fiorino AS. Electron-beam computed tomography, coronary artery calcium, and evaluation of patients with coronary artery disease. *Ann Intern Med* 1998; **128**:839–47. Reason for exclusion: not fully systematic – only details literature search

35. Floriani I, Ciceri M, Torri V, Tinazzi A, Jahn H, Noseda A. Clinical profile of ioversol – a metaanalysis of 57 randomized, double-blind clinical trials. *Invest Radiol* 1996;**31**:479–91. Reason for exclusion: not assessing diagnostic accuracy

36. Flynn K, Adams E, Anerson D. *Positron emission tomography: systematic review*. Boston, MA: Veterans Affairs Medical Center, Health Services Research and Development Service, Management Decision and Research Center; 1996. Reason for exclusion: superseded by Adams, 1998[56]

37. Frishberg BM. The utility of neuroimaging in the evaluation of headache in patients with normal neurologic examinations. *Neurology* 1994; **44**:1191–7. Reason for exclusion: not assessing accuracy

38. Glasziou PP, Woodward AJ, Mahon CM. Mammographic screening trials for women aged under 50. A quality assessment and meta-analysis. *Med J Aust* 1995;**162**:625–9. Reason for exclusion: not assessing diagnostic accuracy

39. Goffinet F, Paris J, Nisand I, Breart G. Utilité clinique du Doppler ombilical. Résultats des essais contrôles en population à haut risque et a bas risque. *J Gynecol Obstet Biol Reprod Paris* 1997; **26**(1):16–26. Reason for exclusion: not assessing diagnostic accuracy

40. Goffinet F, Paris-Llado J, Nisand I, Breart G. Umbilical artery Doppler velocimetry in unselected and low risk pregnancies: a review of randomised controlled trials. *Br J Obstet Gynaecol* 1997;**104**:425–30. Reason for exclusion: not assessing diagnostic accuracy

41. Goodman M, Lamm SH, Engel A, Shepherd CW, Houser OW, Gomez MR. Cortical tuber count: a biomarker indicating neurologic severity of

tuberous sclerosis complex. *J Child Neurol* 1997;**12**:85–90. Reason for exclusion: not assessing diagnostic accuracy

42. Griffiths AM, Sherman PM. Colonoscopic surveillance for cancer in ulcerative colitis: a critical review. *J Pediatr Gastroenterol Nutr* 1997; **24**:202–10. Reason for exclusion: not assessing diagnostic accuracy

43. Guise J, Mahon S, Aickin M, Helfand M, Peipert JF, Westhoff C. Screening for bacterial vaginosis in pregnancy. *Am J Prev Med* 2001; **20** (3 Suppl):1–20. Reason for exclusion: screening; not test accuracy

44. Hailey D, Sampietro-Colom L, Marshall D, Rico R, Granados A, Asua J. The effectiveness of bone mineral density measurement and associated treatments for prevention of fractures: an international collaborative review. *Int J Technol Assess Health Care* 1998;**14**:237–54. Reason for exclusion: not assessing diagnostic accuracy

45. Hailey D, Tomie JA. An assessment of gait analysis in the rehabilitation of children with walking difficulties. *Disabil Rehabil* 2000;**22**:275–80. Reason for exclusion: not diagnostic accuracy

46. Harris KM, Kelly S, Berry E, Hutton J, Roderick P, Cullingworth J, *et al*. Systematic review of endoscopic ultrasound in gastro-oesophageal cancer. *Health Technol Assess* 1998;**2**(18). Reason for exclusion: prognostic review

47. Hearn J, Higginson IJ. Do specialist palliative care teams improve outcomes for cancer patients: a systematic literature review. *Palliat Med* 1998; **12**:317–32. Reason for exclusion: not assessing diagnostic accuracy

48. Helfand M, Redfern CC. Clinical guideline, part 2: screening for thyroid disease: an update. *Ann Intern Med* 1998;**129**:144–58. Reason for exclusion: not assessing diagnostic accuracy

49. Hirtz D, Ashwal S, Berg A, Bettis D, Camfield C, Camfield P, *et al*. Practice parameter: evaluating a first nonfebrile seizure in children: report of the quality standards subcommittee of the American Academy of Neurology, The Child Neurology Society, and The American Epilepsy Society. *Neurology* 2000;**55**:616–23. Reason for exclusion: not diagnostic accuracy – yield of abnormality

50. Homik J, Hailey D. *Quantitative ultrasound for bone density measurement*. Edmonton, AB: Alberta Heritage Foundation for Medical Research; 1998. Reason for exclusion: not assessing diagnostic accuracy

51. Hong MK, Mintz GS, Popma JJ. Limitations of angiography for analyzing coronary atherosclerosis progression or regression. *Ann Intern Med* 1994; **121**:348–54. Reason for exclusion: not assessing diagnostic accuracy

52. Hsu CW, Imperiale TF. Meta-analysis and cost comparison of polyethylene glycol lavage versus sodium phosphate for colonoscopy preparation. *Gastrointest Endosc* 1998;**48**:276–82. Reason for exclusion: not assessing diagnostic accuracy

53. Huijgen HJ, Sanders GTB, Koster RW, Vreeken J, Bossuyt PMM. The clinical value of lactate dehydrogenase in serum: a quantitative review. *Eur J Clin Chem Clin Biochem* 1997;**35**:569–79. Reason for exclusion: did not assess test for diagnosis of specific condition but across several specialties – insufficient detail given

54. Ishida H, Takemura Y, Kawai T. A systematic review for the diagnostic accuracy of serum C-reactive protein measurement in neonatal infants with infection. *Rinsho Byori* 2001; **49**:1020–9. Reason for exclusion: non-English

55. Ivanov RI, Allen J, Sandham JD, Calvin JE. Pulmonary artery catheterization: a narrative and systematic critique of randomized controlled trials and recommendations for the future. *New Horizons* 1997;**5**:268–76. Reason for exclusion: not assessing diagnostic accuracy

56. Jensen LA, Onyskiw JE, Prasad NG. Meta-analysis of arterial oxygen saturation monitoring by pulse oximetry in adults. *Heart Lung* 1998;**27**:387–408. Reason for exclusion: not assessing diagnostic accuracy

57. Jorm AF. Methods of screening for dementia: a meta-analysis of studies comparing an informant questionnaire with a brief cognitive test. *Alzheimer Dis Assoc Disord* 1997;**11**:158–62. Reason for exclusion: not assessing diagnostic accuracy

58. Katon W, Gonzales J. A review of randomized trials of psychiatric consultation–liaison studies in primary care. *Psychosomatics* 1994;**35**:268–78. Reason for exclusion: not assessing diagnostic accuracy

59. Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. *JAMA* 1995;**273**:149–54. Reason for exclusion: not assessing diagnostic accuracy

60. Koger SM, Chapin K, Brotons M. Is music therapy an effective intervention for dementia? A meta-analytic review of literature. *J Music Ther* 1999; **36**:2–15. Reason for exclusion: not assessing diagnostic accuracy

61. Kroenke K, Taylor-Vaisey A, Dietrich AJ, Oxman TE. Interventions to improve provider diagnosis and treatment of mental disorders in primary care. A critical review of the literature. *Psychosomatics* 2000;**41**:39–52. Reason for exclusion: not assessing diagnostic accuracy

62. Kurz X, Kahn SR, Abenhaim L, Clement D, Norgren L, Baccaglini U, *et al*. Chronic venous disorders of the leg: epidemiology, outcomes, diagnosis and management: summary of an

evidence-based report of the VEINES task force. *Int Angiol* 1999;**18**:83–102. Reason for exclusion: not focused on diagnostic accuracy; limited details of methods provided

63. Lachner G, Engel RR. Differentiation of dementia and depression by memory tests. A meta-analysis. *J Nerv Ment Dis* 1994;**182**:34–9. Reason for exclusion: not assessing diagnostic accuracy

64. Lau J, Ioannidis J, Balk E, Milch C, Chew P, Terrin N, *et al*. *Evaluation of technologies for identifying acute cardiac ischemia in emergency departments*. Rockville, MD: Agency for Healthcare Research and Quality; 2001. Reason for exclusion: same report as Lau, 2001[71]

65. Laxson CJ, Titler MG. Drawing coagulation studies from arterial lines: an integrative literature review. *Am J Crit Care* 1994;**3**:16–24. Reason for exclusion: not assessing diagnostic accuracy

66. Linzer M, Yang EH, Estes M, Wang P, Vorperian VR, Kapoor WN. Diagnosing syncope part 1: value of history, physical examination, and electrocardiography. *Ann Intern Med* 1997; **126**:989–96. Reason for exclusion: not diagnostic accuracy – diagnostic yield

67. Lokeshwar V, Soloway M. Current bladder tumor tests: does their projected utility fulfill. *J Urol* 2001;**165**:1067–77. Reason for exclusion: no systematic review methods, other than a literature search, were reported

68. Losier BJ, McGrath PJ, Klein RM. Error patterns on the continuous performance test in non-medicated and medicated samples of children with and without ADHD: a meta-analytic review. *J Child Psychol Psychiatry* 1996;**37**:971–87. Reason for exclusion: not assessing diagnostic accuracy

69. Maenza RL, Seaberg D, D'Amico F. A meta-analysis of blunt cardiac trauma: ending myocardial confusion. *Am J Emerg Med* 1996; **14**:237–41. Reason for exclusion: not assessing diagnostic accuracy

70. Mannerkorpi K, Ekdahl C. Assessment of functional limitation and disability in patients with fibromyalgia. *Scand J Rheumatol* 1997;**26**:4–13. Reason for exclusion: not diagnostic accuracy

71. Markert RJ, Walley ME, Guttman TG, Mehta R. A pooled analysis of the Ottawa ankle rules used on adults in the ED. *Am J Emerg Med* 1998; **16**:564–7. Reason for exclusion: not assessing diagnostic accuracy

72. Marshall D, Johnell O, Wedel H. Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. *BMJ* 1996; **312**:1254–9. Reason for exclusion: not assessing diagnostic accuracy

73. Messori A, Trippoli S, Becagli P, Tendi E. Treatments for newly diagnosed advanced ovarian cancer: analysis of survival data and cost-effectiveness evaluation. *Anticancer Drugs* 1998; **9**:491–502. Reason for exclusion: not assessing diagnostic accuracy

74. Michelson E, Hollrah S. Evaluation of the patient with shortness of breath: an evidence based approach. *Emerg Med Clin North Am* 1999; **17**:221–37. Reason for exclusion: not diagnostic accuracy; reported % of patients with various symptoms

75. Moreyra E, Finkelhor RS, Cebul RD. Limitations of transesophageal echocardiography in the risk assessment of patients before nonanticoagulated cardioversion from atrial fibrillation and flutter: an analysis of pooled trials. *Am Heart J* 1995; **129**:71–5. Reason for exclusion: not assessing diagnostic accuracy

76. Morris PS. A systematic review of clinical research addressing the prevalence, aetiology, diagnosis, prognosis and therapy of otitis media in Australian aboriginal children. *J Paediatr Child Health* 1998; **34**:487–97. Reason for exclusion: not assessing diagnostic accuracy

77. Munro J, Booth A, Nicholl J. Routine preoperative testing: a systematic review of the evidence. *Health Technol Assess* 1997;**1**(12). Reason for exclusion: not assessing diagnostic accuracy

78. Murray J, Cuckle H, Taylor G, Littlewood JO, Hewison J. Screening for cystic fibrosis. *Health Technol Assess* 1999;**3**(8). Reason for exclusion: not assessing diagnostic accuracy

79. Muttreja MR, Mohler ER. Clinical use of ischemic markers and echocardiography in the emergency department. *Echocardiography* 1999;**16**:187–92. Reason for exclusion: not fully systematic – only search and inclusion criteria documented

80. New Zealand Health Technology Assessment. *Colour vision screening: a critical appraisal of the literature*. Christchurch: New Zealand Health Technology Assessment; 1998. Reason for exclusion: not assessing diagnostic accuracy

81. New Zealand Health Technology Assessment. *Screening programmes for the detection of otitis media with effusion and conductive hearing loss in pre-school and new entrant school children: a critical appraisal of the literature*. Christchurch: New Zealand Health Technology Assessment; 1998. Reason for exclusion: not assessing diagnostic accuracy

82. Nwosu CR, Khan KS, Chien PF, Honest MR. Is real-time ultrasonic bladder volume estimation reliable and valid? a systematic overview. *Scand J Urol Nephrol* 1998;**32**:325–30. Reason for exclusion: not assessing diagnostic accuracy

83. Ofman JJ, Rabeneck L. The effectiveness of endoscopy in the management of dyspepsia: a qualitative systematic review. *Am J Med* 1999;

106:335–46. Reason for exclusion: not assessing diagnostic accuracy

84. Oliveira CA, Troster EJ, Pereira CR. Inhaled nitric oxide in the management of persistent pulmonary hypertension of the newborn: a meta-analysis. *Rev Hosp Clin Fac Med Sao Paulo* 2000;**55**:145–54. Reason for exclusion: not assessing diagnostic accuracy

85. Oostveen JC, van de Laar MA. Magnetic resonance imaging in rheumatic disorders of the spine and sacroiliac joints. *Semin Arthritis Rheum* 2000;**30**:52–69. Reason for exclusion: no systematic review methods

86. Ornato JP, Selker HP, Zalenski RJ. Overview: diagnosing acute cardiac ischemia in the emergency department. A report from the National Heart Attack Alert Program. 2001; *Ann Emerg Med* **37**:450–2. Reason for exclusion: background comment

87. Palomaki GE, Neveux LM, Haddow JE. Can reliable Down's syndrome detection rates be determined from prenatal screening intervention trials? *J Med Screen* 1996;**3**:12–17. Reason for exclusion: not assessing diagnostic accuracy

88. Pearson V. *Antenatal ultrasound scanning*. York: University of York, NHS Centre for Reviews and Dissemination; 1994. Reason for exclusion: not assessing diagnostic accuracy

89. Pembrey ME, Barnicoat AJ, Carmichael B, Bobrow M, Turner G. An assessment of screening strategies for fragile X syndrome in the UK. *Health Technol Assess* 2001;**5**(7). Reason for exclusion: not systematic review of accuracy

90. Pietrobon R, Coeytaux RR, Carey TS, Richardson WJ, DeVellis RF. Standard scales for measurement of functional outcome for cervical pain or dysfunction: a systematic review. *Spine* 2002;**27**:515–22. Reason for exclusion: not assessing diagnostic accuracy

91. Pirkis JE, Jolley D, Dunt DR. Recruitment of women by GPs for pap tests: a meta-analysis. *Br J Gen Pract* 1998;**48**:1603–7. Reason for exclusion: not assessing diagnostic accuracy

92. 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). Reason for exclusion: not really about diagnostic accuracy; more about screening policies

93. 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. Reason for exclusion: no systematic review methods reported

94. Ratner P, Bottorff J, Johnson J, Cook R, Lovato C. A meta-analysis of mammography screening promotion. *Cancer Detect Prev* 2001;**25**:147–60. Reason for exclusion: not assessing diagnostic accuracy

95. Richard CS, McLeod RS. Follow-up of patients after resection for colorectal cancer: a position paper of the Canadian Society of Surgical Oncology and the Canadian Society of Colon and Rectal Surgeons. *Can J Surg* 1997;**40**:90–100. Reason for exclusion: not assessing diagnostic accuracy

96. Rimer BK, Bluman LG. The psychosocial consequences of mammography. *J Natl Cancer Inst* 1997;**22**:131–8. Reason for exclusion: not assessing diagnostic accuracy

97. Ringertz H, Marshall D, Johansson C, Johnell O, Kullenberg RJ, Ljunghall S, *et al*. Bone density measurement – a systematic review. A report from SBU, the Swedish Council on Technology Assessment in Health Care. *J Intern Med Suppl* 1997;**241** (Suppl 739):1–60. Reason for exclusion: not assessing diagnostic accuracy

98. Rosenthal M, Christensen BK, Ross TP. Depression following traumatic brain injury. *Arch Phys Med Rehabil* 1998;**79**(1):90–103. Reason for exclusion: not assessing diagnostic accuracy

99. Salekin RT, Rogers R, Sewell KW. A review and meta-analysis of the psychopathy checklist and psychopathy checklist-revised: predictive validity of dangerousness. *Clin Psychol Sci Pract* 1996; **3**:203–15. Reason for exclusion: abstract indicates not accuracy

100. 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. Reason for exclusion: duplicate publication – summary of Ioannidis, 2001[154,155] and Balk, 2001[70]

101. Seymour CA, Cockburn F, Thomason MJ, Littlejohns P, Chalmers RA, Lord J, *et al*. Newborn screening for inborn errors of metabolism: a systematic review. *Health Technol Assess* 1997; **1**(11). Reason for exclusion: not assessing diagnostic accuracy

102. Shaw LJ, Eagle KA, Gersh BJ, Miller DD. Meta-analysis of intravenous dipyridamole–thallium-201 imaging (1985 to 1994) and dobutamine echocardiography (1991 to 1994) for risk stratification before vascular surgery. *J Am Coll Cardiol* 1996;**27**:787–98. Reason for exclusion: not accuracy – evaluates prognostic value of imaging tests, but not compared against a reference test

103. Snowdon SK, Stewart-Brown SL. *Preschool vision screening: results of a systematic review*. York: NHS Centre for Reviews and Dissemination; 1997. Reason for exclusion: not assessing diagnostic accuracy

104. Song JC, White CM. Do HMG-CoA reductase inhibitors affect fibrinogen? *Ann Pharmacother* 2001;**35**:236–41. Reason for exclusion: not assessing diagnostic accuracy

105. Spitalnic SJ, Woolard RH, Mermel LA. The significance of changing needles when inoculating blood cultures. *Clin Infect Dis* 1995;**21**:1103–6. Reason for exclusion: not assessing diagnostic accuracy

106. Storgaard H, Nielsen SD, Gluud C. The validity of the Michigan Alcoholism Screening Test (MAST). *Alcohol Alcohol* 1994;**29**:493–502. Reason for exclusion: not assessing diagnostic accuracy

107. Thacker SB, Stroup DF, Peterson HB. Efficacy and safety of intrapartum electronic fetal monitoring: an update. *Obstet Gynecol* 1995;**86** (4 Part 1): 613–20. Reason for exclusion: not diagnostic accuracy

108. Thijs L, Staessen JA, Celis H, de Gaudemaris R, Imai Y, Julius S, *et al*. Reference values for self-recorded blood pressure: a meta-analysis of summary data. *Arch Intern Med* 1998;**158**:481–8. Reason for exclusion: not assessing diagnostic accuracy

109. Tresch DD. Diagnostic and prognostic value of ambulatory electrographic monitoring in older patients. *J Am Geriatr Soc* 1995;**43**:66–70. Reason for exclusion: not assessing diagnostic accuracy

110. Trpkova B, Major P, Prasad N, Nebbe B. Cephalometric landmarks identification and reproducibility: a meta analysis. *Am J Orthod Dentofacial Orthop* 1997;**112**:165–70. Reason for exclusion: not assessing diagnostic accuracy

111. Turp JC, Minagi S. Palpation of the lateral pterygoid region in TMD – where is the evidence? *J Dent* 2001;**29**:475–83. Reason for exclusion: not assessing diagnostic accuracy

112. van der Wurff P, Hagmeijer RH, Meyne W. Clinical tests of the sacroiliac joint. A systemic methodological review. Part 1: reliability. *Man Ther* 2000;**5**:30–6. Reason for exclusion: evaluates reliability not diagnostic accuracy – see van der Wurff, 2000[249]

113. 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**(4):427–34. Reason for exclusion: not diagnostic accuracy

114. Vasquez TE, Rimkus DS, Hass MG, Larosa DI. Efficacy of morphine sulfate-augmented hepatobiliary imaging in acute cholecystitis. *J Nucl Med Technol* 2000;**28**:153–5. Reason for exclusion: not systematic review – only documents search

115. Vernon SW. Participation in colorectal cancer screening: a review. *J Natl Cancer Inst* 1997; **89**:1406–22. Reason for exclusion: not assessing diagnostic accuracy

116. Wagner TH. The effectiveness of mailed patient reminders on mammography screening: a meta-analysis. *Am J Prevent Med* 1998;**14**:64–70. Reason for exclusion: not assessing diagnostic accuracy

117. Wald NJ, Kennard A, Hackshaw A, McGuire A. Antenatal screening for Down's syndrome. *Health Technol Assess* 1998;**2**(1). Reason for exclusion: not a systematic review; no methods reported

118. Westwood ME, Kelly S, Berry E, Bamford JM, Gough MJ, Airey CM, *et al*. Use of magnetic resonance angiography to select candidates with recently symptomatic carotid stenosis for surgery: systematic review. *BMJ* 2002;**324**:198. Reason for exclusion: duplicate of Berry, 2002[82]

119. White MJ, Nichols CN, Cook RS, Spengler PM, Walker BS, Look KK. Diagnostic overshadowing and mental retardation: a meta-analysis. *Am J Ment Retard* 1995;**100**:293–8. Reason for exclusion: not assessing diagnostic accuracy

120. Wilkie D, Savedra MC, Holzemer WL, Tesler MD, Paul SM. Use of the McGill Pain Questionnaire to measure pain: a meta-analysis. *Nurs Res* 1990; **39**:36–41. Reason for exclusion: not assessing diagnostic accuracy

121. Willems M, Quartero AO, Numans ME. How useful is paracetamol absorption as a marker of gastric emptying? A systematic literature study. *Dig Dis Sci* 2001;**46**:2256–62. Reason for exclusion: not diagnostic accuracy

122. Wise EA. Diagnosing posttraumatic stress disorder with the MMPI clinical scales: a review of the literature. *J Psychopathol Behav Assess* 1996; **18**:71–82. Reason for exclusion: not assessing diagnostic accuracy

123. Yabroff KR, Kerner JF, Mandelblatt JS. Effectiveness of interventions to improve follow-up after abnormal cervical cancer screening. *Prev Med* 2000;**31**:429–39. Reason for exclusion: not assessing diagnostic accuracy

124. Zakzanis KK, Leach L, Freedman M. Structural and functional meta-analytic evidence for fronto-subcortical system deficit in progressive supranuclear palsy. *Brain Cogn* 1998;**38**:283–96. Reason for exclusion: not assessing diagnostic accuracy

Appendix 4

Details of review methods

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Adams, 1998[56]	Head and neck, colorectal, breast, lung/solitary pulmonary nodules cancer and Alzheimer's disease	PET	CT and MRI	MEDLINE Other electronic sources Ref. lists Update of previous systematic review	Language: restricted; English only Quality restrictions: sample description	Haynes, 1995;[61] Authors' own	21	Not clearly reported
Anand, 1998[51]	Deep vein thrombosis	Clinical assessment; compression US, impedance plethysomography	Venography	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	Holleman, 1995[62]	5 for clin. assmt	Not reported
Attia, 1999[63]	Thyroid disease in acutely ill hospitalised patients	Clinical signs and symptoms; sTSH	Biochemical markers (including 2nd- or 3rd-generation sTSH assays) ± clinical features and follow-up after resolution of non-thyroidal illness	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	Authors' own	10 (0)	4734 clin. assmt; 431? sTSH
Bachmann, 1998[64]	Diabetic retinopathy	Direct opthalmoscopy, non-stereoscopic retinal photography	Retinal examination or stereoscopic retinal photography	MEDLINE EMBASE Other electronic sources (SCI) Ref. lists	Language: not stated Quality restrictions: Appropriate ref. test Sample description Adequate description of test	Sackett, 1991[65] as inclusion criteria	9 (2)	Not reported

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Bader, 2001[66]	Dental caries	Visual and visual/tactile inspection, radiography, fibre-optic transillumination, electrical conductance, laser fluorescence alone or in combination	Histology	MEDLINE EMBASE Other electronic sources	Language: restricted; English only Quality restrictions: none	Authors' own	39	?
Badgett, 1997[50]	Left-sided heart failure in adults	Clinical examination	Left ventricular end diastolic pressure, left atrial pressure, pulmonary capillary wedge pressure, pulmonary artery diastolic pressure	MEDLINE	Language: not stated Quality restrictions: appropriate ref. test	Modified Holleman, 1995[62]	34 (0)	21,660
Badgett, 1996[67]	Left ventricular dysfunction	Chest radiographic findings (cardiomegaly, redistribution/congestion, interstitial oedema)	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	MEDLINE Ref. lists Other sources	Language: restricted; English only Quality restrictions: appropriate ref. test	Authors' own	29 (?)	Not reported
Bafounta, 2001[68]	Melanoma	Dermoscopy (also known as epiluminescence microscopy) vs naked eye	Histology	MEDLINE EMBASE Other electronic sources: PASCAL, BIOMED and BIUM Ref. lists	Language: no restriction Quality restrictions: appropriate ref. test, blinding used, sample description	Cite Irwig, 1994;[6] Cochrane Methods Working Group, 1996[69]	8 (8)	2193

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Balk, 2001[70] (methods in Lau, 2001[71])	Acute cardiac ischaemia (AMI or unstable angina)	Biochemical markers	Not clear – WHO definition for AMI?	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: none	Authors' own? reference Irwig, 1994[6]	77? (7)	?
Banks, 2001[72]	Breast cancer	Breast cancer screening with use of HRT	Not reported, presumably follow-up	MEDLINE Other electronic sources – SCI Ref. lists	Language: no restriction Quality restrictions: none	Not conducted	8 (7 accuracy)	
Barton, 1999[73]	Breast cancer	CBE	Follow-up with or without mammography	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: appropriate ref. test	Not conducted	6 (0)	Not reported
Bastian, 1998[74]	Pregnancy	Home pregnancy test kits	Laboratory-based urine or serum hCG test	MEDLINE Other electronic sources Ref. lists Other sources	Language: restricted; English only Quality restrictions: appropriate ref. test	Holleman, 1995[62] – as inclusion criteria	5	Approx. 1449
Bastian, 1997[75]	Early pregnancy	Clinical examination (history, symptoms), home pregnancy tests	Detection of B subunit of hCG in urine or serum	MEDLINE Ref. lists	Language: restricted; English language only Quality restrictions: appropriate ref. test	Holleman, 1995[62]	9	Approx. 6445
Becker, 1996[76]	Acute venous thromboembolism	D-dimer testing	Venography, pulmonary angiography, ventilation–perfusion scanning, lower-extremity ultrasonography	MEDLINE Other electronic sources Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test, sample description	Becker, 1989[77]	29	4200

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Bell, 1998[78]	Ovarian cancer	Ultrasound, CA 125	Histological examination of ovarian tissue for positive results and follow-up for negative	MEDLINE Other electronic sources Ref. lists Contact experts	Language: not stated Quality restrictions: none	Cochrane Methods Working Group checklist[69]	25	Median 2572
Berger, 2000[79]	Gallstones	Abdominal symptoms (upper abdominal pain, biliary colic, radiating pain, use of analgesics, tenderness of upper abdomen, food intolerance, fat intolerance)	US or oral cholecystography	MEDLINE Ref. lists	Language: restricted; English, French, Dutch, German Quality restrictions: appropriate ref. test	Authors' own	24 (0)	36302
Berry, 1999[80]	Hepatic lesions, pulmonary embolus and CAD	Spiral and EBCT	Angiography	MEDLINE EMBASE Other electronic sources Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test	Authors' own (Kelly, 2004[81])	49 (7 in meta-analysis for EBCT, 0 paired)	1246 in meta-analysis
Berry, 2002;[82] Westwood, 2002[83]	Carotid artery disease (stenosis or tandem lesions); peripheral vascular disease (severity)	MRA (contrast-enhanced, 3D TOF, 2D TOF, phase-contrast)	Intra-arterial DSA; cut-film angiography (same as X-ray angiography?)	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts Internet	Language: no restriction Quality restrictions: avoidance of verification bias	Authors' own	30 (3)	NR
Bjelland, 2002[84]	Anxiety disorders and depression	HADS	Structured or semi-structured diagnostic interview, e.g. DSM-III	MEDLINE Other electronic sources: ISI, PsycINFO	Language: not stated Quality restrictions: appropriate ref. test	Not conducted	24 (accuracy)	4620

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Blakeley, 1995[85]	Carotid artery stenosis	Non-invasive carotid artery tests including: carotid Doppler ultrasonography; real-time B-mode US; duplex US; MRA; supraorbital Doppler US; oculoplethysmography	Carotid angiography or intra-arterial carotid DSA	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test	Modified Sackett, 1991[65]	70 (?)	6406
Bonis, 1997[86]	Hepatitis C (to identify prognostic benefit)	Biochemical tests	Histological outcome (liver biopsy?)	MEDLINE Ref. lists	Language: no restriction Quality restrictions: avoidance of VB	Mulrow, 1989[87]	42; 15 pooled (0)	773 (pooled)
Bradley, 1998[88]	Heavy drinking and/or alcohol abuse in women	Alcohol screening questionnaires with ≤ 10 items	In-depth interviews based on standard criteria	MEDLINE Other electronic sources Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test	Authors' own	9	>10476
Buchanan, 2001[89]	Violence in people with dangerous severe personality disorders	Clinical judgement; statistically derived rating of dangerousness	Occurrence of violent episode (usually proxy measures such as arrest, hospital admission or conviction)	MEDLINE Other electronic sources	Language: not stated Quality restrictions: none	Not conducted	21 (?)	13572
Buntinx, 1997[90]	Urological cancer	Macroscopic haematuria	Not described – presumably histology?	MEDLINE Other electronic sources Ref. lists	Language: not stated Quality restrictions: avoidance of VB consecutive enrolment	Authors' own	20 (0)	3161

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Cabana, 1995[91]	Acute cholecystitis	MA-HBS; C-HBS	Surgery with confirmation by pathology, autopsy or clinical follow-up with establishment of an alternative diagnosis	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: appropriate ref. test	Authors' own (as inclusion criteria)	9 (0)	214 C-HBS, 166 MA-HBS
Campens, 1997[92]	Renal function (glomerular filtration rate)	Kidney function tests: serum creatinine, serum urea, 24-hour creatinine clearance, creatinine clearance calculated using the Cockcroft–Gault formula, urea clearance and (urea clearance + creatinine clearance)/2	Insulin clearance, 125-iothalamte clearance, ^{51}Cr-EDTA (chromium edetic acid), clearance, ^{99m}Tc-DTPA (diethylenetriamine-pentaacetic acid) clearance	MEDLINE Other electronic sources Ref. lists Contact experts	Language: restricted; English, German, Dutch, French Quality restrictions: none	Authors' own	26 (4)	7504
Carlson, 1994[93]	Ovarian cancer	Pelvic imaging, US or CA 125 radioimmunoassay (CA 125) as a screening device	Operative confirmation	MEDLINE Ref. lists Authors' own files	Language: restricted; English only Quality restrictions: none	Not conducted	34 (0)	Ultrasound, 12,115; CA 125, 30,555
Cher, 2001[94]	Silicone breast implant rupture	MRI	Surgical removal of implant	MEDLINE Ref. lists	Language: not stated Quality restrictions: none	Authors' own	18 (0)	1039
Chesson, 1997[95]	Wide range of sleep disorders	Polysomnography, oximetry, full or less than full respiratory recording	Attended polysomnography	MEDLINE	Language: restricted; English only Quality restrictions: none	No formal assessment		

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Chien, 1997[96]	Preterm delivery	Cervico-vaginal fetal fibronectin	Preterm delivery before 37 or 34 weeks of gestation and delivery within 1 week after testing	MEDLINE Ref. lists	Language: no restriction Quality restrictions: none	Authors' own adapted from Dunn, 1995;[97] Guyatt, 1992[98] and Cochrane Methods Working Group 1996[69]	14 (0)	723; 847
Choi, 2001[99]	Idiopathic vasculitides (WG, MPA, CSS, iNCGN	Antineutrophil cytoplasmic antibody (pANCA) tests (specifically those including assays for anti-MPO antibodies)	Definitions of the Chapel Hill consensus conference (for WG, MPA and CSS), iNCGN was defined as iNCGN without features of systemic diseases	MEDLINE Ref. list	Language: restricted; English only Quality restrictions: appropriate ref. test, consecutive enrolment	Not conducted	7 (6)	4261
Clarke, 2000[100]	Idiopathic Parkinson's disease	Acute levodopa and apomorphine challenge tests	Clinical diagnosis?	MEDLINE Other electronic sources Ref. lists Contact experts	Language: no restriction Quality restrictions: none	No validity assessment described	13 (3)	645
Conde-Agudelo, 1998[101]	Down syndrome	Triple marker test	Follow-up on pregnancy outcome	MEDLINE Ref. lists	Language: restricted; English, French, German Quality restrictions: none	Authors' own	20 (0)	194,326

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Cuzick, 1999[102]	Cervical abnormalities	Human papilloma virus (HPV) testing (Southern blotting, dot blot, filter *in situ* hybridisation, *in situ* hybridisation, hybrid capture, PCR)	Histology	MEDLINE EMBASE Other electronic sources Ref. lists Other sources	Language: restricted; English only Quality restrictions: none	No formal assessment	Not clearly reported	
Da Silva, 1995[103]	Neonatal septicaemia	Leukocyte indices and C-reactive protein	Bacterial culture from blood, CSF or urine; or positive culture of body fluids obtained from normally sterile sites post-mortem histopathological diagnosis of meningitis or pneumonia	MEDLINE EMBASE Ref. lists	Language: no restriction Quality restrictions: appropriate ref. test	Authors' own	16	2219
D'Arcy, 2000[52]	Carpal tunnel syndrome	History taking and physical examination	Electrocardiographic testing	MEDLINE Ref. lists	Language: restricted Quality restrictions: appropriate ref. test, blinding used	Not conducted	12 (5?)	Can't tell
De Bernardinis, 1999[104]	Acute pancreatitis (prediction of severity and/or prognosis)	Ranson's prognostic signs	Radiological, clinical, surgical, CAT, ECHO, post-mortem and ERCP	MEDLINE Other electronic sources Ref. lists	Language: not stated Quality restrictions: consecutive enrolment	Authors' own	23 (0)	Severity prediction, 2796; prognosis, 1513
de Bruyn, 2001[105]	Cirrhosis	Physical examination	Liver biopsy	MEDLINE Ref. lists Contact experts Other sources	Language: not stated Quality restrictions: none	Authors' own	12 (3)	1895

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
de Vries, 1996[106]	Peripheral arterial disease	Duplex and colour-guided duplex US	Contrast angiography	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	Authors' own	14 (0)	613
Deville, 2000[107]	Source of lower back pain (detection of disc hernias)	The test of Lasegue (SLR/CSLR)	Surgery	MEDLINE EMBASE Ref. lists	Language: not stated Quality restrictions: appropriate ref. test	Cochrane Methods Working Group[69]	15 (8)	7395
Devous, 1998[108]	Epilepsy	SPECT brain imaging	EEG or surgical outcome	MEDLINE: EMBASE Other electronic sources Ref. lists	Language: not stated Quality restrictions: none	Authors' own	30 (?)	465??
Dharnidharka, 2002[109]	Kidney function	Serum cystatin C; serum creatinine	GFR, i.e. clearance of inulin or the tracers Cr-EDTA, Tm-DTPA, iothalamate or iohexol	Search not clearly reported	Language: not stated Quality restrictions: none	Not conducted	11 (?)	997
Di Fabio, 1996[110]	Peripheral vestibular deficits, Menière's disease, benign paroxysmal positional vertigo, central nervous system vestibular-impairment	Platform posturography	Other vestibular tests: electronystagmo-graphy or rotation testing	MEDLINE: Other electronic sources Ref. lists	Language: not stated Quality restrictions: appropriate ref. test	Not conducted	9 (4?)	1477
Dijkhuizen, 2000[213]	Endometrial carcinoma and hyperplasia	Endometrial sampling (incl Pipelle, Endo-pap, Pistolet, Accurette)	D&C, hysteroscopy and/or hysterectomy	MEDLINE Ref. lists	Language: not stated Quality restrictions: none	Authors' own	39 (2)	7914

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Dinnes, 2001[111]	Breast cancer	Double or single reading of breast screening mammograms	Follow-up	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts	Language: not stated Quality restrictions: none		10 (3 on accuracy)	588,456
Divakaran, 2001[112]	Fetal anaemia	Non-invasive techniques (fetal US and Doppler blood flow velocity)	Fetal haemoglobin	MEDLINE EMBASE Ref. lists	Language: not stated Quality restrictions: appropriate ref. test, prospective only	Authors' own	8	362
Ebell, 2000[113]	AMI	Troponin T and troponin I	WHO criteria or similar	MEDLINE	Language: restricted; English, French, German, Spanish Quality restrictions: appropriate ref. test, prospective only, blinding used	Authors' own (mainly as inclusion criteria)	19 (1)	5794
Eberhard-Gran, 2001[114]	Postnatal depression	EPDS	Clinical interview (e.g. DSM-III)	MEDLINE Ref. lists Science Citation Index Expanded (ISI)	Language: not stated Quality restrictions: none	None reported	18	1678 (interviewed)
ECRI, 1999[115]	Detection of pneumonia or chronic aspiration in people with swallowing disorders (dysphagia)	Full BSE, 3-oz water test, modified barium swallow, fibre-optic endoscopy, other instrumented examinations	VFSS	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts	Language: restricted; English only Quality restrictions: none	Not formally assessed	Not clearly reported (3)	

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Eden, 2001[116]	Thyroid cancer	History and physical examination (palpation of thyroid gland); FNA	Ultrasound; findings at surgery	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: none	Not described	7 on palpation vs ultrasound; 15 on FNA vs surgery	4616 (palpation; not reported for FNA)
Eiberg, 2001[117]	Peripheral occlusive arterial disease of the lower limb	MRA (TOF or CE)	CA, IOA or IAP	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test, prospective only	Not conducted	28 (4?)	572 TOF-MRA, 387 CE-MRA
Ernst, 1999[118]	Diagnosis of any medical condition	Iridology	Various	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts Other sources	Language: no restriction Quality restrictions: blinding used	Not conducted	4	385
Fahey, 1995[119]	Cervical precancer	Pap test	Histology	MEDLINE Ref. lists Contact experts	Language: not stated Quality restrictions: appropriate ref. test, unpublished studies excluded	Authors' own	62 (0)	17,421
Faron, 1998[120]	Prediction of preterm delivery	Fetal fibronectin	Follow-up	MEDLINE Other electronic sources Ref. lists Contact experts Other sources: Current Contents, Index Medicus conference proceedings	Language: no restriction Quality restrictions: prospective only, blinding used, complete follow-up	Authors' own (as inclusion criteria)	29 (0)	8159

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Fiellin, 2000[121]	Alcohol problems	Screening methods	Identified diagnostic instrument or operational definition (e.g. quantity and frequency of alcohol consumption)	MEDLINE only	Language: restricted; English only Quality restrictions: appropriate ref. test	Authors' own	38	
Fiorino, 1998[122]	CAD	EBCT to detect coronary artery calcium	Angiography	MEDLINE Other electronic sources Ref. lists	Language: restricted; English only Quality restrictions: prospective only	None reported	14 on EBCT	3301
Fischer, 2001[123]	Lung cancer	PET and gamma-camera PET	CT	MEDLINE EMBASE Other electronic sources Ref. lists	Language: restricted; English, German, French Quality restrictions: none	Adams, 1998[56]	55 (0)	>800 for dedicated PET; >400 for gamma-camera PET
Fleischmann, 1998[124]	CAD	Exercise echocardiography; exercise SPECT imaging	Coronary angiography	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: appropriate ref. test	Authors' own	24; 27 (6)	2637; 3237
Fowlie, 1998[125]	Bacterial infection from birth to 90 days	Haematological indices; C reactive protein evaluation; surface swab assessment	Diagnosis of infection (criteria required described)	MEDLINE	Language: restricted; English only Quality restrictions: none	Authors' own	194	

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Frost, 1999[126]	Acute or chronic ankle instability	Stress radiography (anterior drawer or talar tilt)	Surgical confirmation	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test	Not conducted	8	950
Garzon, 2001[127]	Restenosis after percutaneous transluminal coronary angioplasty	ETT; stress nuclear imaging; stress echocardiographic imaging	Coronary angiography at 6 months (quantitative or semi-quantitative assessment)	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test, blinding used	Authors' own (as inclusion criteria)	13 (5)	Various (see results)
Gianrossi, 1990[128]	CAD	Cardiac fluoroscopy	Coronary angiography	MEDLINE Other electronic sources Ref. list	Language: no restriction Quality restrictions: none	Wachter, 1988[129]	13 (0)	3765
Gifford, 2000[130]	Evaluation of dementia	Clinical prediction rules for neuroimaging	CT or MRI	MEDLINE Ref. lists	Language: not stated Quality restrictions: avoidance of VB	None reported	7	1505
Gottlieb, 1999[131]	Deep venous thrombosis in symptomatic patients	Calf sonography	Contrast venography	MEDLINE Ref. lists Other sources	Language: restricted; English only Quality restrictions: appropriate ref. test, blinding used, sample description	Authors' own (as inclusion criteria)	? (0)	
Gould, 2001[132]	Pulmonary nodules and mass lesions	PET	Histology	MEDLINE Other electronic sources Contact experts	Language: no restriction Quality restrictions: none	Adapted Kent, 1992[133]	40 (0)	1909

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Gronseth, 2000[134]	Multiple sclerosis	Evoked potential identified clinically silent lesions	Follow-up	MEDLINE Ref. lists	Language: not stated Quality restrictions: appropriate ref. test	Authors' own	9	892
Hallan, 1997[135]	Acute appendicitis	C-reactive protein; total leucocyte count	Histology	MEDLINE Other electronic sources Ref. lists Other sources	Language: not stated Quality restrictions: appropriate ref. test	Authors' own	22 (13)	3436
Harvey, 1996[136]	Depression in stroke patients	DST	Clinical examination	MEDLINE Other electronic sources Ref. lists	Language: not stated Quality restrictions: none	None reported	9	352
Heffner, 1995[137]	Complicated parapneumonia effusions that require drainage	Pleural fluid pH, LDH and glucose	Combination of diagnostic test results and determinations of patient outcome	MEDLINE Ref. lists	Language: not stated Quality restrictions: none	Irwig, 1994[6]	7	274
Heffner, 1997[138]	Exudative and transudative pleural effusions	Biochemical pleural fluid tests: P-R (pleural fluid to serum ratio), LDH-PF (lactate dehydrogenase to pleural fluid ratio), LDH-R (LDH to serum ratio), P-PF (protein to pleural fluid ratio), C-PF (pleural to serum creatinine ratio), C-R (pleural fluid to serum ratio), A-G (pleural fluid to serum albumin gradients), BILI-R (pleural fluid to serum bilirubin ratios)	Clinical assessment – using explicit, objective and reproducible criteria beyond clinical judgment alone/including positive biopsy specimens	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	Authors' own	7	1448

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Helfand, 2001[139]	Skin cancer including melanoma	Screening tests (total-body skin examination, partial skin examination, lesion-specific examination and others)	Skin biopsy	MEDLINE Ref. lists Contact experts	Language: not stated Quality restrictions: none	Not conducted	5 accuracy	13,636
Hider, 1999[140]	Early breast cancer	Clinical examination, FNA, triple test, core biopsy, ultrasound, scintigraphy, magnetic resonance mammography, breast self-examination	Mammography; histology	MEDLINE EMBASE Other electronic sources Ref. lists	Language: restricted; English only Quality restrictions: none	New Zealand Health Technology Assessment Centre[141]	Not clearly reported	
Hobbs, 1997[142]	Various	Near-patient testing	Laboratory-based methods	MEDLINE EMBASE Other electronic sources – Science Citation Index, GP-Lit, CINAHL Ref. lists Contact experts Other sources	Language: not stated Quality restrictions: none	Reid, 1995[143]	32	Not reported
Hobby, 2001[144]	Tears of the triangular fibrocartilage complex, the intrinsic carpal ligaments and osteonecrosis of the carpal bones	MRI imaging of the wrist	Arthroscopy	MEDLINE EMBASE Ref. lists Other sources	Language: restricted; English only Quality restrictions: appropriate ref. test	None reported	16 (0)	956

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Hoffman, 2000[145]	Prostate cancer in men with non-specific PSA levels (4.0–10.0 ng/ml)	Free-to-total PSA ratio	Multiple systematic transrectal prostate needle biopsy or long-term follow-up if negative, radical prostatectomy, transurethral resections	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	Authors' own	17 pooled (0)	3123 in MA
Hoffman, 2002[146]	Prostate cancer	Free-to-total PSA ratio	Biopsy	MEDLINE Ref. lists Other sources	Language: restricted; English only Quality restrictions: none	Not conducted	12	
Hofman, 2000[147]	ICH in patients with mild head injury	Plain skull radiograph or CT	CT, follow-up to determine uneventful recovery, angiography, neurosurgical findings	MEDLINE EMBASE Other electronic sources – Current Contents Ref. lists	Language: not stated Quality restrictions: none	None reported	13 (0)	53,494
Hooft, 2001[148]	Recurrent papillary or follicular thyroid carcinoma	[18F] Fluorodeoxyglucose PET	Histology/cytology, focal ^{131}I uptake, pathognomic bone scan or MRI for bone metastases, CT/MRI for brain metastases, progression of radiologically documented lesions suspect for malignancy	MEDLINE EMBASE Other electronic sources – CancerLit, CDSR Ref. lists	Language: no restriction Quality restrictions: none	Cochrane Methods Working Group checklist[69]	14	402

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Hrung, 1999[149]	Breast cancer	MRI	Excisional biopsy or mastectomy	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: appropriate ref. test	Authors' own	16 (0)	1494
Huicho, 2002[150]	Urinary tract infection in children	Urine screening tests [leucocyturia or pyuria in centrifuged urine; bacteria and/or leucocytes in uncentrifuged, stained or unstained urine; and dipstick tests (LE and nitrite)]	Quantitative urine culture	MEDLINE Other electronic sources Ref. lists Contact experts	Language: restricted; English, Spanish Quality restrictions: appropriate ref. test, avoidance of VB	Modified Mulrow, 1989[87]	48 ('several')	31,458 samples
Huicho, 1996[151]	Acute infectious diarrhoea	Faecal screening tests (including combination of faecal leucocytes with clinical data)	Stool culture	MEDLINE Ref. lists Contact experts	Language: no restriction Quality restrictions: appropriate ref. test, avoidance of VB	Mulrow, 1989[87]	25 (5)	19,036
Hurley, 2000[152]	Detection of Gram-negative bacteraemia in patients with suspected bacterium	2 types of LAL assay: GLAL and CLAL	Blood culture	MEDLINE Ref. lists Contact experts	Language: not stated Quality restrictions: appropriate ref. test	None reported	56 (2?)	4134
Ioannidis, 2001[153]	Acute uncomplicated sinusitis in children	Diagnostic methods including plain radiograph, CT, ultrasound, sinus aspirate	Clinical diagnosis, plain radiograph, sinus aspirate	MEDLINE	Language: restricted; English only Quality restrictions: none	None reported	21 (8 accuracy)	1524 accuracy

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Ioannidis, 2001[154] (methods in Lau, 2001[71])	Acute cardiac ischaemia	Electrocardiography (out-of-hospital)	WHO definition for AMI	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: none	Irwig, 1994[6]	11; 8 pooled (0)	7508
Ioannidis, 2001[155] (methods in Lau, 2001[71]	Acute cardiac ischaemia	Echocardiography; technetium-99m sestamibi scanning	WHO definition for AMI	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: none	Irwig, 1994[6]	16 (1)	Not clear
Kallmes, 1996[156]	Carotid artery stenosis	MRA as confirmatory test following positive sonography	Conventional angiography (film screen or digital subtraction)	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test	Not conducted	17 (0)	Total not reported
Kearon, 1998[157]	Deep venous thrombosis	Non-invasive approaches: impedence plethysmography and venous US alone or in combination with other tests	Venography	MEDLINE Ref. lists Other sources	Language: not stated Quality restrictions: appropriate ref. test, prospective only, blinding used, consecutive enrolment	Authors' own – used as inclusion criteria	50 (0)	
Kim, 2001[158]	Coronary disease	Pharmacological stress testing (dipyridamole SPECT imaging or echocardiography)	Vasodilator or inotropic agent combined with single-photon emission CT imaging or echocardiography	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: none	Irwig, 1994[6]	82 (11)	7995

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Kinkel, 2000[159]	Ovarian masses	US characterisation – morphological assessment, Doppler US, color Doppler, Doppler flow imaging	Histopathological findings	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test, blinding used	None reported	46 (?)	5159
Kinkel, 1999[160]	Endometrial cancer	Radiological staging (including CT, US and MRI	Surgical staging with histopathological results	MEDLINE Ref. lists	Language: restricted; English, Japanese, Italian, French, German Quality restrictions: appropriate ref. test, blinding used	Authors' own	47 (5)	1779
Kittler, 2002[161]	Cutaneous melanoma	Dermoscopy	Histopathology	MEDLINE Ref. lists Contact experts	Language: restricted; English or German Quality restrictions: none	Authors' own	27 (3?)	9821
Klompas, 2002[162]	Acute thoracic aortic dissection	Clinical history, physical examination, plain chest X-ray	Surgical exploration, autopsy, aortogram, MRI, CT, transoesophageal echocardiography	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test, consecutive enrolment	Authors' own	21 (10)	1848
Knopman, 2001[163]	Dementia	Clinical diagnostic criteria	Neuropathological confirmation	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts	Language: restricted; English only Quality restrictions: none	Existing tool for classification of evidence (not referenced)	Not clear	Not clear

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Koelemay, 1996[164]	Peripheral arterial occlusive disease	Duplex US	Angiography	MEDLINE Ref. lists	Language: restricted; English, German, Dutch Quality restrictions: appropriate ref. test	Authors' own	16 (0)	Not clearly reported
Koelemay, 2001[165]	Lower extremity arterial disease	MRA	Conventional arteriography or intra-arterial digital subtraction angiography	MEDLINE EMBASE Other electronic sources Ref. lists	Language: restricted; English, German, French Quality restrictions: none	Authors' own	34 (3)	1090
Koumans, 1998[53]	*Neisseria gonorrhoeae*	Nucleic acid hybridisation test (Pace 2 or Pace 2C); nucleic acid amplification tests [ligase chain reaction (LCR) or polymerase chain reaction (PCR)]	Culture	MEDLINE Ref. lists Other sources	Language: not stated Quality restrictions: none	Authors' own	21 (0)	17,737 (Pace 2, 13,236; LCR, 4501)
Kowalski, 2001[166]	HIV	Enzyme immunosorbent assays	Western blot, PCR, p24 antigen testing	MEDLINE Other electronic sources – AIDSLINE Other sources	Language: restricted; English only Quality restrictions: none	Adapted Cooper, 1988[167]	16 (?)	Diseased, 3–1654; non-diseased, 2–4999
Kwok, 1999[168]	CAD in women	Exercise tests: exercise ECG, exercise radionuclide scan, exercise echo	Coronary angiography	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: appropriate ref. test	Authors' own	21 (0)	

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Lacasse, 1999[169]	Localised pulmonary lesions	Transthoracic needle aspiration biopsy	Resection specimen, biopsy procedures of an adjacent site with tumour involvement, long-term follow-up or culture	MEDLINE Ref. lists Other sources	Language: restricted; English only Quality restrictions: appropriate ref. test, avoidance of VB, consecutive enrolment	Authors' own	48 (0)	9047 biopsies
Lau, 1999;[170] Engels, 2000[171]	Acute bacterial rhinosinusitis	Clinical features and imaging technologies (sinus radiography, US, MRI, endoscopy or CT)	Sinus puncture with bacterial culture; investigators' diagnoses	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts Other sources	Language: described as restricted to English only but 'several studies published in other languages were included in the EPCs analyses' Quality restrictions: avoidance of VB	Authors' own	14 (5)	Approx. 2424
Law, 1998[172]	Speech and language delay	Screening tests	Norm-referenced tests or objectified clinical judgements	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts Other sources	Language: not stated Quality restrictions: appropriate ref. test	Authors' own	45 (85 data sets, 19 = 1 test)	Not clearly reported
Lederle, 1999[173]	Asymptomatic AAA	Physical examination (abdominal palpation)	Ultrasound	MEDLINE Ref. lists Contact experts Other sources	Language: no restriction Quality restrictions: none	Holleman, 1995[62]	15 (0)	2955

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Leitich, 1999[174]	Preterm delivery	Cervicovaginal fetal fibronectin	Follow-up (timing of delivery)	MEDLINE EMBASE Ref. lists	Language: restricted; English only Quality restrictions: prospective only	Not conducted	28 comparisons (0)	Not clear
Li, 2001[175]	Endotracheal tube placement confirmation	End-tidal capnography	Not reported (confirmation by separate standard required)	MEDLINE Other electronic sources – NIH database Ref. lists Contact experts Other sources	Language: no restriction Quality restrictions: none	Not conducted	10 (0)	2192
Liedberg, 1996[176]	Temperomandibular joint disorder	Arthrography, CT, MRI	Surgery, clinical + imaging, cryosection, macroscopy, arthrography	MEDLINE Ref. lists Other sources	Language: restricted; English only Quality restrictions: none	Authors' own	31 (0)	1185 joints
Lindbaek, 2002[177]	Acute purulent sinusitis	Clinical examination	Sinus puncture, CT, X-ray, US	MEDLINE	Language: restricted; English only Quality restrictions: appropriate ref. test, prospective only	Cochrane Collaboration Methods group[69]	4	1016
Littenberg, 1995[178]	Low back pain	SPECT bone imaging	Surgical results or long-term follow-up	MEDLINE EMBASE Other electronic sources Biological Abstracts Ref. lists	Language: restricted; English only Quality restrictions: Appropriate ref. test	Authors' own	6	

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Loy, 1996[179]	*Helicobacter pylori* infection	Commercial serological kits (ELISA and latex agglutination)	Culture, histology, urase testing on biopsy and other	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: none	Adapted Jaeschke, 1994[39,180]	21 (8)	Not reported
Lysakowski, 2001[181]	Vasospasm in patients with subarachnoid haemorrhage due to ruptured aneurysm	TCD	Cerebral angiography	MEDLINE EMBASE Other electronic sources Ref. lists	Language: no restriction Quality restrictions: appropriate ref. test	Adapted Lijmer, 1999[15]	26 (0)	Not reported
Mackenzie, 1996[182]	Meniscal and cruciate ligament disorders	MRI of the knee	Arthroscopy	MEDLINE EMBASE Other electronic sources Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test	Authors' own	22 (2)	2929
Mango, 1998[183]	Cervical cancer	INNA cervical cancer screening	Biopsy, independent pathologist or site pathologist interventions	MEDLINE Manufacturer's database	Language: not stated Quality restrictions: none	Authors' own	22 (0)	214,590
Markert, 1998[184]	Fractured ankle or foot	Ottawa ankle rules (original and revised)	Ankle and foot radiography	MEDLINE 'Computerised and manual literature searches'	Language: Not stated Quality restrictions: appropriate ref. test	Not conducted	7 (2)	4213
Mayer, 1997[185]	Malignant melanoma	Dermatoscopy and clinical diagnosis (also known as dermoscopy, skin surface microscopy, epiluminescence microscopy and incident light microscopy)	Excision biopsy with histopathological examination	MEDLINE EMBASE Ref. lists	Language: no restriction Quality restrictions: appropriate ref. test	Sackett, 1991[65] criteria used	6	1382

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
McCrory, 1999[186]	Abnormal cervical cytology	Pap smear; newer cytological methods	Histology, histology or negative colposcopy or cytology	MEDLINE EMBASE Other electronic sources	Language: not stated Quality restrictions: appropriate ref. test	Authors' own	109 (0?)	
McGee, 1999[187]	Hypovolaemia in adults	Physical/bedside diagnosis	Serum urea nitrogen–creatinine ratio; plasma osmolarity, serum osmolality or serum sodium, % weight gain after rehydration, hypotension or postural pulse increment	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	Authors' own?	4 clinical studies	179
McNaughton-Collins, 2000[188]	Chronic abacterial prostatitis	Culture for detection of infection; zinc levels; ultrasound	Meares–Stamey, prostate histology	MEDLINE Other electronic sources Ref. lists	Language: not stated Quality restrictions: none	Reid, 1995[143]	4	Various
Merritt, 1997[189]	Cervical metastasis	Physical examination and CT	Not described	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	Not conducted	12 (12)	667
Metlay, 1997[190]	Community acquired pneumonia	History and physical examination	New infiltrate on chest radiograph	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	Authors' own	4	Not reported

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Mitchell, 1999[191]	Squamous intraepithelial lesions of the cervix	Fluorescence spectroscopy, Pap smear, cervicography, speculoscopy, HPV testing	Colposcopic-directed biopsy	MEDLINE Ref. lists Other sources	Language: not reported Quality restrictions: appropriate ref. test	Not conducted	58 (2)	
Mitchell, 1998[192]	Squamous intraepithelial lesions of the cervix	Colposcopy	Colposcopic-directed biopsy	MEDLINE Ref. lists Other sources	Language: not reported Quality restrictions: appropriate ref. test	Not conducted	9 (0)	6281
Mol, 1999[193]	Down syndrome	Nuchal translucency measurement	Fetal karyotype	MEDLINE EMBASE Ref. lists	Language: not stated Quality restrictions: appropriate ref. test, consecutive enrolment	Authors' own	25 (0)	67,990
Mol, 1998[44]	Ectopic pregnancy; pregnancy failure	Single serum progesterone measurement	Various: surgery, histology, sonography, delivery, D&C, dropping hCG	MEDLINE EMBASE Ref. lists	Language: not reported Quality restrictions: none	Authors' own	26 (0)	8926
Mol, 1997[194]	Tubal pathology in subfertile patients	Chlamydia antibody titres	Laparoscopy	MEDLINE EMBASE Ref. lists	Language: not stated Quality restrictions: appropriate ref. test	Authors' own	23 (1)	2729
Mol, 1998[195]	Endometriosis	CA-125 serum	Laporoscopy	MEDLINE EMBASE Ref. lists	Language: not reported Quality restrictions: appropriate ref. test	Authors' own	30 (0)	2866

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Mol, 1998[196]	Predicting successful IVF	SPA	Results of IVF	MEDLINE	Language: not stated Quality restrictions: none	None reported	24 (0)	647
MSAC, 1999[197]	Permanent congenital hearing impairment	OAEA	Other form of audiology (auditory brainstem response, visual reinforcement audiometry, distraction, play audiometry)	MEDLINE EMBASE Other electronic sources Ref. lists Other sources	Language: restricted; English only Quality restrictions: none	Modified Irwig, 1994[6]	12	2233
Mullins, 2000[198]	Pulmonary embolism	Spiral volumetric CT	Pulmonary arteriogram or another clinical reference standard, e.g. V/Q scan	MEDLINE Other electronic sources Ref. lists	Language: restricted; English only Quality restrictions: none	Authors' own	11	762
Muris, 1994[199]	Dyspepsia (organic cause)	Clinical signs and symptoms	Endoscopy, ultrasound	MEDLINE Other electronic sources Ref. lists	Language: restricted; English only Quality restrictions: none	Authors' own	10	7224
Muris, 1992[200]	Colorectal cancer	Clinical diagnosis	Endoscopy or long-term follow-up	MEDLINE Ref. lists	Language: not stated Quality restrictions: none	Authors' own	14	3308
Mushlin, 1998[201]	Breast cancer	Mammography (with or without CBE)	Histology, follow-up and referral for further work-up (for specificity)	MEDLINE Ref. lists	Language: not stated Quality restrictions: none	Not conducted	9 (0)	263,359

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Mustafa, 2002[202]	Upper extremity deep vein thrombosis	US	Venography	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: prospective only	Jaeschke, 1994[39,180]	6	170
Nallamothu, 2001[203]	CAD	EBCT	Coronary angiography	MEDLINE Other electronic sources Ref. lists Contact experts	Language: restricted; English only Quality restrictions: appropriate ref. test	Authors' own	14 (0)	1662
Nanda, 2000[204]	Cervical cytological abnormalities	Papanicolau test (conventional methods, computer screening or rescreening or monolayer cytology)	Histological examination, colposcopy, cytology	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts	Language: restricted; English only Quality restrictions: appropriate ref. test	Authors' own	97 (0)	
Nelson, 2001[54]	Chlamydial genitourinary infections	Screening tests including DNA amplification tests	Culture	MEDLINE Other electronic sources Ref. lists Contact experts	Language: restricted; English only Quality restrictions: sample description, test appropriately performed, all tests appropriately used, sufficient information on sample selection	Authors' own (as inclusion criteria)	? Impossible to say	
Nuovo, 1997[205]	Cervical cancer	Cervicography	Colposcopy with or without directed biopsies	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: avoidance of VB	Authors' own	7	1773

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Oei, 1995[206]	Pregnancy in infertile couples/failure to conceive	Post-coital test	Follow-up	MEDLINE Ref. lists	Language: not stated Quality restrictions: avoidance of VB	Not conducted	11 (0)	4007
Olatidoye, 1998[207]	Prognosis of unstable angina pectoris	Troponin T or troponin I	Follow-up for non-fatal MI or cardiac death	MEDLINE Ref. lists	Language: no restriction Quality restrictions: none	Not conducted	20 (1)	2847; 1901
Oosterhuis, 2000[208]	Vitamin B_{12} deficiency	MCV assessed using microbiological assays or immunological assays	Low serum vitamin B_{12} concentration with or without additional diagnostic investigations	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	Authors' own	37 (0)	1121
Orr, 1995[209]	Appendicitis	US	Surgery (only in those with positive US)	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	QA was not performed	17 (0)	3358
Owens, 1996[55]	HIV in adults	PCR	Enzyme immunoassay with confirmatory western blot; viral culture; antigen testing	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts Other sources	Language: restricted; English only Quality restrictions: none	Authors' own	96 (0)	14,668

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Owens, 1996[210]	HIV in infants	PCR	Viral culture; persistence of HIV antibody past 15 months; definitive clinical evidence of HIV (Center for Disease Control system). Sustained loss of HIV antibody considered evidence of absence of infection	MEDLINE Other electronic sources – searched 17 databases up to 1991; MEDLINE alone for 1992–1994 Ref. lists	Language: restricted; English only Quality restrictions: none	Authors' own	32 (0)	1796
Pasternack, 2001[211]	Humeral epicondylitis	MRI	Clinical diagnosis	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts	Language: not stated Quality restrictions: none	Authors' own	7	148
Patel, 2000[212]	Vaginal trichomoniasis	PCR, ELISA, direct fluorescence antibody test, different culture media	Trichomonads culture in one or more media with/without the wet mount	MEDLINE Ref. lists	Language: restricted – unable to translate papers in Slovak, Polish, Russian, Korean or Czech Quality restrictions: none	Irwig, 1994[6]	35 (2?)	9882
Pearl, 1996[214]	Blunt abdominal trauma	US	DPL, CT or laporotomy	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: prospective only	Kent, 1992[133]	11	2000

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Peters, 1996[215]	Diabetes	Glycosylated haemoglobin (HbA_{1c})	OGTT	MEDLINE Ref. lists Other sources	Language: articles with English abstracts Quality restrictions: none	None conducted	10 (0)	8984
Petersen, 2001[216]	Early detection of dementia	Screening instruments: Mini-Mental State Examination and general screening instruments; Clock Drawing and Time Change tests; neuropsychological batteries; informant-based instruments	Not described, 'independent standard for dementia'	MEDLINE EMBASE Other electronic sources – Current Contents, Psychological Abstracts, PsycInfo, Cochrane Database, CINAHL Ref. lists	Language: restricted; English only? Quality restrictions: none	Classification of evidence	74	Not clear
Rao, 1995[217]	Wegener granulomatosis	Antineutrophil cytoplasmic antibody (c-ANCA)	Standard reference criteria (e.g. Ear, Nose, Throat, Lung and Kidney staging system, Fauci criteria and the American College of Rheumatology criteria – first and last require biopsy confirmation)	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test, consecutive enrolment	Authors' own	15 (0)	13,562
Rao, 1999[218]	Type 2 diabetes	Risk factors for diabetes (aim was to preselect high-risk patients who should undergo serum screening)	Various: presence of diabetes (with follow-up for prospective studies or retrospective identification); blood glucose levels	MEDLINE EMBASE Other electronic sources – Cochrane database Ref. lists	Language: restricted; English only Quality restrictions: none	Jaeschke 1994[39,180]	7	8289 (6 studies; total not reported for 1 study)

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Rao, 1999[219]	Spinal canal compromise and cord compression in cervical spinal cord injury	Cervical radiography, myelograms, CT myelogram, CT, MRI, ultrasound	Not clear	MEDLINE Ref. lists	Language: restricted; English, German, French Quality restrictions: none	Not conducted	37	Not reported
Rathbun, 2000[220]	Pulmonary embolism	Helical CT	Pulmonary angiography, or normal ventilation–perfusion lung scan for absence of pulmonary embolism	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: prospective only	Jaeschke, 1994[39,180]	15	1330
Reed, 1996[221]	Community-acquired pneumococcal pneumonia	Sputum Gram stain	Culture	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: blinding used	Author's own	12 (0)	Not reported
Revah, 1998[222]	Preterm birth	Fetal fibronectin	Follow-up	MEDLINE	Language: restricted; English only Quality restrictions: prospective only, blinding used	Not conducted	24 (0)	2737 symptomatic; 4042 asymptomatic
Ross, 1999[223]	Sleep apnoea	Sleep monitoring devices, radiological imaging, laboratory assays, clinical signs and symptoms	Sleep laboratory polysomnogram-derived apnoea index, apnoea-hypoapnoea index or respiratory distress index	MEDLINE Other electronic sources Ref. lists Other sources	Language: restricted; English, German, French, Spanish, Italian Quality restrictions: none	Authors' own	147 (?)	17,679

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Safriel, 2002[224]	Pulmonary emboli	CT pulmonary angiography	Fluoroscopic pulmonary angiography, scintigraphy	MEDLINE Ref. lists Other sources	Language: restricted; English only Quality restrictions: none	Not conducted	12 (0)	1250
Scheid, 2001[225]	Nephropathy in patients with diabetes	Quantitative or semiquantitative screening tests for microalbuminuria	Not reported	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	McKibbon, 1995[226]	176 (?)	9824
Scheidler, 1997[227]	Lymph node metastasis in patients with cervical cancer	LAG, CT, MRI	Histological or cytological specimens obtained by surgery or lymph node biopsy	MEDLINE Ref. lists Contact experts	Language: restricted; English only Quality restrictions: appropriate ref. test, blinding used	Not conducted	38 (5)	
Schwimmer, 2000[228]	Recurrent cutaneous melanoma	2-Fluro-2-deoxy-D-glucose positron emission tomography FDG-PET	Not reported (could be surgery, biopsy, follow-up)	MEDLINE Other electronic sources	Language: not stated Quality restrictions: none	Authors' own	13; 8 pooled (0)	982 (not clearly reported)
Scouller, 2000[229]	Excessive alcohol consumption	CDT, GGT	Alcohol consumption, diagnosis of abuse or dependence or other relevant criteria for assessing drinking	MEDLINE Other electronic sources	Language: restricted; English only Quality restrictions: none	Authors' own	110 (37)	Not reported
Selley, 1997[230]	Prostate cancer	Digital rectal examination, prostate specific antigen, transrectal ultrasound imaging, colour Doppler imaging	Biopsy and histology	MEDLINE EMBASE Ref. lists Other sources	Language: no restriction Quality restrictions: none	Not conducted	Not clearly reported	

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Siegman-Igra, 1997[231]	Vascular catheter-related bloodstream infection	Laboratory diagnostic methods (catheter segment culture, blood culture of blood drawn through catheter, etc.)	Clinical definition of catheter-related bloodstream infection or catheter-related infection or catheter segment culture	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: none	Not conducted	28; 22 pooled (4 max.)	
Smith, 1996[232]	Trisomy 21 (as marker for fetus with Down syndrome)	Triple-screen (α-fetoprotein, hCG, unconjugated estriol) vs double-screen α-fetoprotein and hCG)	Down syndrome babies born or aborted	MEDLINE	Language: restricted; English only Quality restrictions: prospective only	Not conducted	8 (6)	143,094
Smith-Bindman, 2001[233]	Down syndrome	Ultrasound (second trimester)	Chromosomal analysis or visual inspection	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test, avoidance of VB, complete follow-up, sample description	Authors' own (as inclusion criteria)	56 (0)	132,295
Smith-Bindman, 1998[234]	Endometrial cancer (exclusion of)	Endovaginal ultrasound	Histology/biopsy	MEDLINE Ref. lists	Language: no restriction Quality restrictions: prospective only, blinding used	None	35 (0)	5892
Solomon, 2001[235]	Torn meniscus or ligament of the knee	Physical examination	Arthroscopy, arthrotomy, MRI	MEDLINE Other electronic sources Ref. lists	Language: restricted; English only Quality restrictions: none	Holleman, 1995[62]	23 (0)	Not clearly reported

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Sonnad, 2001[238]	Prostate cancer staging	MRI	Pathology	MEDLINE	Language: restricted; English only Quality restrictions: none	Single criterion used	27 (4)	1796
Spencer-Green, 1997[237]	Systemic sclerosis	Antibody tests (anti-centromere and anti-Scle-70)	Authors' classification of systemic sclerosis (previously published)	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test	Mulrow, 1989[87]	30 (?)	Not reported; >1000
Stengel, 2001[238]	Blunt abdominal trauma	US	CT, surgery, peritoneal lavage, clinical observation	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts	Language: no restriction Quality restrictions: none	Authors' own plus Sackett 2000[239]	30 (0)	9047
Storgaard, 1994[240]	Alcohol problems	MAST	Other defined diagnostic criteria of alcohol problems	MEDLINE Ref. lists	Language: not stated Quality restrictions: none	Not conducted	20	4433
Swart, 1995[241]	Tubal pathology (absence of tubal patency and presence of peritubal adhesions)	HSG	Laporoscopy with chromopertubation	MEDLINE Ref. lists	Language: restricted; English, French, German, Dutch Quality restrictions: none	Authors' own	20 (0)	4179

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Taylor-Weetman, 2002[242]	Occlusal and proximal dental caries in the posterior teeth of the deciduous and permanent dentitions	Panoramic radiography (OPT) and bitewing radiography	Histology (for extracted teeth) or a suitable reference standard for UK practice should be standard panoramic and bitewing radiographs with a clinical examination and follow-up where possible	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts Other sources	Language: no restriction Quality restrictions: none	Authors' own	5	12,806 teeth or surfaces
Tenner, 1994[243]	Gallstone pancreatitis	Laboratory tests including bili (bilirubin), AD (alkaline phosphatase), ALT (alanine aminotransferase), AST (aspartate transaminase)	Sonography, CT, ERCP, laporotomy	MEDLINE	Language: restricted; English only Quality restrictions: none	Not conducted	8 (4)	557
Helfand, 2001[244]	Hearing loss	Newborn hearing screening using OAEs or ABR	Behavioural test (e.g. visual reinforcement audiometry)	MEDLINE Other electronic sources Ref. lists Contact experts	Language: restricted; English only Quality restrictions: none	US Preventive Services Task Force[245] Quality Rating	11	
Tugwell, 1997[246]	Lyme disease	Laboratory diagnosis, including culture, ELISA, western blot	Criteria for confirmed infection	MEDLINE Other sources	Language: restricted; English only Quality restrictions: blinding used, sample description, reproducible information on sampling and reference standard, diagnosis by experts blinded to test results	Irwig, 1994[6] (as inclusion criteria)	8 on various tests	Not clearly reported

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
van Beek, 2001[247]	Pulmonary embolism	Lung scintigraphy; helical CT	Pulmonary angiography; lung scintigraphy	MEDLINE Other electronic sources	Language: restricted Quality restrictions: appropriate ref. test, prospective only, blinding used, consecutive enrolment, sample description	Authors' own (used as inclusion criteria)	15 scintigraphy; 12 spiral CT (0)	Not reported; 1171
van den Hoogen, 1995[248]	Low back pain resulting from radiculopathy, vertebral cancer metastasis, ankylosing spondylitis	History, physical examination, and erythrocyte sedimentation rate	Anatomical findings at surgery; overall clinical impression after diagnostic imagery or New York Criteria (for ankylosing spondylitis)	MEDLINE Ref. lists Other sources	Language: not stated Quality restrictions: none	Author's own	36	Not reported
van der Wurff, 2000[249]	Sacro-iliac joint	Clinical tests – pain provocation or mobility	Various – primarily visual analogue scale	MEDLINE EMBASE Other electronic sources Ref. lists	Language: restricted; English, French, Dutch, German Quality restrictions: none	Authors' own (see van der Wurff, 2000[250])	11	Not clearly reported
Varonen, 2000[251]	Acute maxillary sinusitis	Ultrasound, radiography, clinical examination	Sinus puncture or CT	MEDLINE Other electronic sources Ref. lists Contact experts	Language: restricted; English, German, French, Scandinavian languages, Finnish Quality restrictions: appropriate ref. test	Cochrane Methods Working Group, 1996[69]	9 (6)	1144

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Vasbinder, 2001[252]	Renal artery stenosis	Computed tomography angiography, MRA, US, captopril renal scintigraphy, captopril test	Intra-arterial X-ray angiography	MEDLINE EMBASE Other electronic sources Ref. lists	Language: restricted; English, German, French Quality restrictions: appropriate ref. test, avoidance of VB	Authors' own	55 (10)	
Verkooijen, 2000[253]	Non-palpable breast disease	Large-core needle biopsy	Surgical biopsy or minimum of 2 years' follow-up in at least 90% of patients	MEDLINE Ref. lists	Language: restricted; English only Quality restrictions: appropriate ref. test	Not formally conducted	5 (0)	483?
Visser, 2000[254]	Peripheral arterial disease (evaluation of arterial stenoses and occlusions)	Gadolinium-enhanced MRA; colour-guided duplex ultrasound	Conventional angiography	MEDLINE Ref. lists Contact experts Other sources	Language: no restriction Quality restrictions: appropriate ref. test	Authors' own plus Kent, 1992[133] (latter not described)	9 MRA; 18 US	216 MRA; 1059 US
Vroomen, 1999[255]	Sciatica due to disc herniation	History and physical examination	Surgery, myelography, CT, MRI	MEDLINE Other electronic sources Ref. lists Contact experts	Language: restricted; English, French, German, Dutch Quality restrictions: none	Sackett, 1991[65]	37 (0)	Not reported
Watson, 2002[256]	Chlamydia	PCR, LCR, gene probe, enzyme immunoassay, direct immunofluorescence, culture	Culture or 'expanded gold standard' (i.e. two non-culture techniques)	MEDLINE EMBASE Other electronic sources Ref. lists Contact experts Other sources	Language: not stated Quality restrictions: none	Irwig, 1994[6]	30 (5)	30,988

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Wells, 1995[257]	Deep vein thrombosis	Ultrasound (real-time B-mode, duplex, colour Doppler)	Standard contrast venography	MEDLINE Other electronic sources Ref. lists	Language: restricted; English only Quality restrictions: avoidance of VB	Authors' own	16 (0)	2001
White, 2000[258]	Intracranial aneurysms	Non-invasive imaging: CT, angiography, MRA and transcranial Doppler US	Interarterial digital subtraction angiography result	MEDLINE EMBASE Ref. lists Other sources	Language: no restriction Quality restrictions: appropriate ref. test	Authors' own (as inclusion criteria)	38 (6)	1765
Whited, 1998[259]	Melanoma	Clinical examination (checklists; global assessment)	Histopathological examination of excised tissue	MEDLINE Other electronic sources Ref. lists	Language: not stated Quality restrictions: appropriate ref. test, avoidance of VB, blinding used	Holleman, 1995[62] (used as inclusion criteria)	12	Not reported
Whitsel, 2000[260]	Autonomic failure in people with diabetes	QTc – heart rate corrected QT interval	Cardiovascular reflex test	MEDLINE Other electronic sources Contact experts	Language: no restriction Quality restrictions: appropriate ref. test	Authors' own	17 (0)	4584
Wiese, 2000[261]	Vaginal trichomoniasis	Pap smear; wet mount	Culture	MEDLINE Ref. lists	Language: restricted; could not translate Slovak, Polish, Russian, Korean or Czech Quality restrictions: appropriate ref. test	Irwig, 1994[6]	30 (7)	9501

continued

Study	Target disorder	Index test(s)	Reference test(s)	Search strategy	Language/quality restrictions	Validity assessment	No. of accuracy studies (no. with paired data)	No. of patients
Wijnberger, 2001[262]	Prediction of neonatal respiratory distress syndrome or hyaline membrane disease	Lamellar body count and lecithin/ sphingomyelin ratio	Clinical, radiographic and therapeutic definition	MEDLINE Ref. lists	Language: not stated Quality restrictions: none	Authors' own	6 (0)	600
Williams, 2002[263]	Clinical depression	Questionnaires and clinical examinations	Standard interviews, such as Structured Clinical Interview, to make a criterion-based diagnosis	MEDLINE Other electronic sources – specialised registry of depression trials	Language: restricted; English only Quality restrictions: none	Authors' own	28 case-finding studies; 14 reliability studies (?)	8639 (case-finding); 1389 (reliability)

AAA, abdominal aortic aneurysm; ABR, auditory brainstem response; AMI, acute myocardial infarction; BSE, bedside examination; CA, contrast arteriography; CA 125, cancer antigen 125; CAD, coronary artery disease; CAT, computed axial tomography; CBE, clinical breast examination; CDT, carbohydrate-deficient transferrin; CE, contrast-enhanced; C-HBS, conventional hepatobiliary scintigraphy; CLAL, chromogenic limilus amoebocyte lystae; CSF, cerebrospinal fluid; CSLR, cross straight leg raising test; CSS, Churg–Strauss syndrome; CT, computed tomography; D&C, dilation and curettage; DPL, diagnostic peritoneal lavage; DSA, digital subtraction angiography; DST, dexamethasone suppression test; EBCT, electron-beam computed tomography; ECG, electrocardiogram; ECHO, echotomography; EEG, electroencephalogram; ELISA, enzyme-linked immunosorbent assay; EPDS, Edinburgh Postnatal Depression Scale; ERCP, endoscopin retrograde cholangiopancreatography; ETT, exercise treadmill testing; FNA, fine needle aspiration; GFR, glomerular filtration rate; GGT, γ-glutamyltransferase; GLAL, gelation limilus amoebocyte lystae; HADS, Hospital Anxiety and Depression Scale; hCG, human chorionic gonadotrophin; HRT, hormone replacement therapy; HSG, hysterosalpingography; IAP, intra-arterial pressure; ICH, intracranial haemorrhage; iNCGN, isolated pauci-immune necrotising and crescentic glomerulonephritis; INNA, interactive neural network-assisted; IOA, intraoperative arteriography; IVF, *in vitro* fertilisation; LAG, lymphangiography; LAL, limilus amebocyte lystae; LDH, lactate deydrogenase; MA-HBS, morphine-augmented hepatobiliary scintigraphy; MAST, Michigan Alcoholism Screening Test; MCV, mean corpuscular volume; MI, myocardial infarction; MPA, microscopic polyangiitis; MPO, myeloperoxidase; MRA, magnetic resonance angiography; OAE, otoacoustic emission; OAEA, oto-acoustic emission audiometry; OGTT, oral glucose tolerance test; OPT, orthopantomogram; PCR, polymerase chain reaction; PET, positron emission tomography; PSA, prostate-specific antigen; QA, quality assurance; SLR, straight leg raising test; SPA, sperm penetration assay; sTSH, sensitive thyrotropin test; TCD, transcranial Doppler; TOF, time-of-flight; TRUS, transrectal ultrasound imaging; US, ultrasonography; VFSS, video fluoroscopy; WG, Wegener's granulomatosis.

Appendix 5

Details of review synthesis methods

Study	Statistical tests			Method of study synthesis					If SROC give means of data presentation				For paired data: separate analysis?	Heterogeneity investigation: method
	Heterogeneity	Threshold effect	Graphical plot	Narrative	Pooled Se/Sp	Other pooled estimate	IPD	SROC	DOR	AUC	Q*	Other		
Adams[56]				✓										Narrative
Anand[51]				✓										Narrative
Attia[63]	χ^2					LR								Narrative
Bachmann[64]			Forest		✓								No	Subgroup
Bader[66]				✓										Narrative
Badgett[50]					✓	LR								Subgroup
Badgett[67]	Statistical test – NR				✓									Subgroup
Bafounta[68]			ROC					Unweighted	✓			LR from Q*	All paired	Subgroup
Balk[70,71]			ROC		✓			Unweighted				Graph only	No	SROC regression Subgroup
Banks[72]	EWLS[a]			✓										None
Barton[73]					✓	LR								Narrative
Bastian[74]	Statistical test – NR				✓Se	ES[b]							No	Subgroup
Bastian[75]	Breslow–Day			✓										None
Becker[76]			ROC	✓										Narrative
Bell[78]				✓										Narrative
Berger[79]	χ^2		ROC			DOR[c]								Log regression
Berry[80]			ROC *D* vs *S*					Unweighted Robust-resistant				TPR at mean FPR		SROC regression
Berry[82,83]			ROC					Weighting NS			✓		No	Subgroup; SROC regression
Bjelland[84]				✓										Narrative
Blakeley[85]	χ^2				✓	ES[b]								Subgroup
Bonis[86]	χ^2		ROC		✓	NPV		Inverse variance weighted				*Se* at fixed *Sp*		Narrative
Bradley[88]				✓										Narrative
Buchanan[89]	Statistical test				✓	Log PV ES[b]								SROC regression
Buntinx[90]	χ^2				✓	PPV								Subgroup
Cabana[91]	χ^2 (Pearson)				✓									None
Campens[92]	χ^2		ROC		✓	PV Correl. coeff.							Yes	Subgroup
Carlson[93]					✓									Subgroup

continued

Study	Statistical tests			Method of study synthesis					If SROC give means of data presentation				For paired data: separate analysis?	Heterogeneity investigation: method
	Heterogeneity	Threshold effect	Graphical plot	Narrative	Pooled Se/Sp	Other pooled estimate	IPD	SROC	DOR	AUC	Q*	Other		
Cher[94]	*Q* statistic for ORs		ROC, CIs		✓	DOR[c]		Weighted				Present graph only		Meta regression
Chesson[95]				✓										Narrative
Chien[96]	Breslow-Day	Spearman				LR								Subgroup
Choi[99]	χ^2		ROC		✓			Samp size weighted				SROC parameters	No	SROC regression
Clarke[100]	Fleiss; comparison of observed vs predicted results		Forest		✓	'Fraction positive'							OR for misclassification	Log regression
Conde-Agudelo[101]	Breslow–Day		ROC		✓ median	PPV		Weighted				*Se* at 95% *Sp*		Subgroup
Cuzick[102]				✓										Narrative
Da Silva[103]	χ^2 or Fishers			✓										Narrative
D'Arcy[52]	ES				✓	LR							No	Method not described
De Bernardinis[104]	χ^2		Funnel plot – ES ROC			ES[b]		Weighted		✓		*Se* at mean *Sp*		SROC regression
de Bruyn[105]	χ^2	Spearman	ROC		✓			Unweighted			✓		No	Subgroup
de Vries[106]	Observed vs predicted CIs for *D*		*D* vs *S*					Inverse variance weighted				*Se* at false positive rate 0.01–0.10 Distance between regression lines		SROC regression
Deville[107]	χ^2; DS&L test for DORs				✓	DOR							No	Meta regression; sensitivity analyses
Devous[108]	DS+L *Q* statistic		Forest *Se*		✓									Log regression; sensitivity analysis
Dharnidharka[109]			AUC per study			AUC								Subgroup
Di Fabio[110]					✓	Effect size[b]							Bonferroni correction	Subgroup ANOVA
Dijkhuizen[213]		Spearman	ROC		✓								No	Subgroup
Dinnes[111]	Narrative statement			✓										Narrative

continued

Study	Statistical tests			Method of study synthesis					If SROC give means of data presentation				For paired data: separate analysis?	Heterogeneity investigation: method
	Heterogeneity	Threshold effect	Graphical plot	Narrative	Pooled Se/Sp	Other pooled estimate	IPD	SROC	DOR	AUC	Q*	Other		
Divakaran[112]	Narrative statement			✓										Narrative
Ebell[113]	Fixed vs random effect results				✓	LR estimated from pooled *Se*/*Sp*		Inverse variance weighted		✓			No	Subgroup
Eberhard-Gran[114]				✓										Narrative
ECRI[115]			ROC					Weighting NS				*Se*/*Sp* at mean threshold	No	Narrative
Eden[116]				✓										None
Eiberg[117]					✓								No	Subgroup
Ernst[118]				✓										
Fahey[119]		Correl. coeff.	ROC					Samp size weighted Unweighted				*Se* as function of FPR SROC parameters		Subgroup; SROC regression
Faron[120]	χ^2				✓	LR								Subgroup
Fiellin[121]				✓										Narrative
Fiorino[122]				✓										Narrative
Fischer[123]					✓	LR								Subgroup
Fleischmann[124]	Predicted vs observed values		ROC		✓			Inverse variance weighted				SROC parameters for 'best/'worst' cases	Yes	Subgroup; SROC regression
Fowlie[125]	Narrative statement			✓										Narrative
Frost[126]				✓										Narrative
Garzon[127]					✓	LR/PV estimated from pooled *Se*/*Sp*							No	Narrative
Gianrossi[128]			*Se* vs prevalence	✓									Subgroup; linear	regression
Gifford[130]	χ^2 or Fisher		ROC including CIs	✓										Narrative
Gottlieb[131]	χ^2				✓	Accuracy								None
Gould[132]			Forest ROC		✓	Log DOR		Weighting NS			✓	*Se* at median *Sp*		Subgroup; sensitivity analysis
Gronseth[134]				✓										Narrative

continued

Study	Statistical tests			Method of study synthesis					If SROC give means of data presentation				For paired data: separate analysis?	Heterogeneity investigation: method
	Heterogeneity	Threshold effect	Graphical plot	Narrative	Pooled Se/Sp	Other pooled estimate	IPD	SROC	DOR	AUC	Q*	Other		
Hallan[135]			ROC *D* vs *S*		✓			Unweighted				SROC parameters[d]	Yes	Subgroup
Harvey[136]				✓										Narrative
Heffner[137]							✓							Narrative
Heffner[138]							✓							None
Helfand[139]				✓										Narrative
Hider[140]				✓										Narrative
Hobbs[142]				✓										Narrative
Hobby[144]	Statistical test – NR				✓	accuracy								Subgroup
Hoffman[145]			ROC including CIs		✓	LR, DOR, AUC		Weighted				*Sp* at *Se* >90% *Se* at *Sp* >90%		Subgroup
Hoffman[146]						Median LR, AUC		Weighted		✓			AUC at single vs multiple cutpoints	
Hofman[147]		Goodness of fit of SROC Spearman on *D*	ROC		✓			Weighting NS				*Sp* at mean *Se*		SROC regression
Hooft[148]				✓										Narrative
Hrung[149]			ROC			LR		Weighting NS			✓	*Sp* at 95% *Se*		Subgroup
Huicho[150]								Weighted				No summary measure	No	SROC regression
Huicho[151]			ROC					Inverse variance weighted				SROC parameters	No	Narrative
Hurley[152]			ROC					Weighting NS			✓		No	Subgroup; sensitivity analyses
Ioannidis[153]	χ^2			✓										Narrative
Ioannidis[71,154]			ROC		✓			Unweighted				Graph only		Subgroup
Ioannidis[155]			ROC		✓			Unweighted Weighted		✓			No	Subgroup
Kallmes[156]					✓									Narrative
Kearon[157]	χ^2				✓	PV								Subgroup
Kim[158]					✓			Inverse variance weighted				Graph only	No	SROC regression Subgroup

continued

Study	Statistical tests			Method of study synthesis					If SROC give means of data presentation				For paired data: separate analysis?	Heterogeneity investigation: method
	Heterogeneity	Threshold effect	Graphical plot	Narrative	Pooled Se/Sp	Other pooled estimate	IPD	SROC	DOR	AUC	Q*	Other		
Kinkel[159]		Pearson	*Se*/*Sp* as function of prev					Robust regression			✓		No	SROC regression Covariate adjustment analysis
Kinkel[160]	Covariate adjustment analysis	Pearson	ROC					Inverse variance weighted			✓		No	SROC regression
Kittler[161]			ROC					Weighting NS[d]	✓log				Yes	SROC regression
Klompas[162]			LR		✓								No	Narrative
Knopman[163]				✓										None
Koelemay[164]	χ^2 or Fishers	Spearman			✓									Subgroup
Koelemay[165]	Breslow–Day for ORs		ROC					Inverse variance weighted			✓	Relative DOR between tests	Yes	SROC regression
Koumans[53]	Assumed homogeneity				✓									Subgroup
Kowalski[166]			Forest		✓								?No	GEE regression
Kwok[168]		Pearson	ROC		✓	LR		Samp size weighted				LRs (method unclear)		SROC regression; sensitivity analysis
Lacasse[169]	χ^2		ROC		✓	LR from pooled *Se*/*Sp*								Subgroup
Lau[170,171]			ROC		✓			Inverse variance weighted		✓			No	Narrative
Law[172]					✓	LR							?No	Subgroup; SROC regression
Lederle[173]	χ^2				✓	LR ES[b]								Subgroup
Leitich[174]	χ^2				✓									Subgroup
Li[175]			Forest		✓									Narrative
Liedberg[176]					✓	LR PV								Subgroup
Lindbaek[177]				✓										Narrative
Littenberg[178]				✓										Narrative
Loy[179]			ROC		✓[f]			Unweighted				*Sp* at mean *Se*	Yes	SROC regression
Lysakowski[181]					✓	LR								Subgroup
Mackenzie[182]					✓	Accuracy							No	Subgroup
Mango[183]					✓									Narrative

continued

Study	Statistical tests			Method of study synthesis					If SROC give means of data presentation				For paired data: separate analysis?	Heterogeneity investigation: method
	Heterogeneity	Threshold effect	Graphical plot	Narrative	Pooled Se/Sp	Other pooled estimate	IPD	SROC	DOR	AUC	Q*	Other		
Markert[184]					✓	PV							No	Subgroup
Mayer[185]				✓										Narrative
McCrory[186]	Narrative statement				✓	ES[b]		Estimated from ES						Log regression on ES
McGee[187]				✓										Narrative
McNaughton-Collins[188]	Narrative statement			✓										Narrative
Merritt[189]	Narrative statement				✓	Accuracy							Yes (all paired)	Subgroup
Metlay[190]				✓										Narrative
Mitchell[191]			ROC		✓			Weighting NS		✓	✓		No	None
Mitchell[192]			ROC		✓			Weighting NS		✓				None
Mol[193]			ROC *Se/Sp* vs sample size		✓	DOR[c]								Log regression; subgroup
Mol[44]	χ^2	Spearman	ROC					Weighting NS				*Se/Sp* from curve		Log regression; subgroup
Mol[194]	χ^2	Spearman	ROC					Weighting NS				*Se/Sp* from curve	No	Log regression; subgroup
Mol[195]	χ^2	Spearman	ROC					Weighting NS				*Se* at fixed *Sp* *Sp* at fixed *Se*		Log regression; subgroup
Mol[196]	χ^2	Spearman	ROC		✓			Weighting NS				*Se* at 95% *Sp*		Subgroup
MSAC[197]				✓										Narrative
Mullins[198]				✓										Narrative
Muris[199]				✓										Narrative
Muris[200]				✓										Narrative
Mushlin[201]	χ^2	Kardaun–Kardaun χ^2 test	ROC					Unweighted				*Se/Sp* from curve		Subgroup; sensitivity analysis
Mustafa[202]				✓										Narrative
Nallamothu[203]	Predicted vs observed values	Spearman	ROC		✓			Variance weighted			✓			SROC regression
Nanda[204]					✓median									Subgroup
Nelson[54]				✓										Narrative
Nuovo[205]				✓										Narrative

continued

Study	Statistical tests			Method of study synthesis					If SROC give means of data presentation				For paired data: separate analysis?	Heterogeneity investigation: method
	Heterogeneity	Threshold effect	Graphical plot	Narrative	Pooled Se/Sp	Other pooled estimate	IPD	SROC	DOR	AUC	Q*	Other		
Oei[206]			ROC			LR								Narrative
Olatidoye[207]			ROC		✓	DOR		Weighted				Graphical comparison of curves	No	None
Oosterhuis[208]	χ^2				✓*Se*									Subgroup; log regression on *Se*
Orr[209]	Fisher	Spearman	Forest *Se*/*Sp*		✓									Subgroup (categorical variables); linear regression (continuous variables)
Owens[55]			Forest			DOR		Estimated from DOR[e]			✓	*Se*/*Sp* read from curve		Subgroup
Owens[210]			Forest		Median			As Owens[55]			✓			Subgroup
Pasternack[211]				✓										Narrative
Patel[212]	χ^2				✓								No	None
Pearl[214]				✓										Narrative
Peters[215]					✓		✓							SROC regression
Petersen[216]				✓										Narrative
Rao[217]	χ^2				✓									Subgroup
Rao[218]				✓										Narrative
Rao[219]				✓										None
Rathbun[220]				✓										Narrative
Reed[221]	Narrative statement		ROC					Samp size weighted				Graph only		SROC regression
Revah[222]					✓	PV								Subgroup
Ross[223]			ROC		✓			Samp size weighted				Graph only	Yes	Narrative
Safriel[224]	Narrative statement				✓			Samp size weighted		✓				Method not described
Scheid[225]	χ^2				✓									Subgroup
Scheidler[227]		Spearman				LR		Weighting NS			✓	LRs from *Q**	No	Subgroup
Schwimmer[228]					✓									Narrative
Scouller[229]			ROC					Unweighted	✓			Ratio of ORs	Yes	SROC regression

continued

Study	Statistical tests			Method of study synthesis					If SROC give means of data presentation				For paired data: separate analysis?	Heterogeneity investigation: method
	Heterogeneity	Threshold effect	Graphical plot	Narrative	Pooled Se/Sp	Other pooled estimate	IPD	SROC	DOR	AUC	Q*	Other		
												(paired studies)		
Selley[230]				✓										Narrative
Siegman-Igra[231]	χ^2 or Fishers; Breslow–Day; Kardaun–Kardaun		ROC		✓	Youden index[f]		Weighting NS				Mean *D*	No	SROC regression
Smith[232]					✓								No	None
Smith-Bindman[233]	Goodness of fit	Spearman	Forest ROC		✓	LR								Subgroup; log regression for *Se*/*Sp*
Smith-Bindman[234]	CI overlap	Correlation	Forest		✓	LR		Weighting NS				Compared curves between tests		Subgroup
Solomon[235]					✓	LR								
Sonnad[236]			ROC					Weighting NS			✓		No	Subgroup; SROC regression
Spencer-Green[237]			*Se* per study		✓								No	Subgroup
Stengel[248]	χ^2		ROC NPV vs Pr			LR		Inverse variance weighted			✓			SROC regression; subgroup
Storgaard[240]				✓										Narrative
Swart[241]	χ^2	Spearman	ROC including CIs		✓			Weighting NS				Graph only		Subgroup
Taylor-Weetman[242]				✓										Narrative
Tenner[243]					✓	PV		Method not clear					No	None
Helfand[244]				✓										None
Tugwell[246]					✓	LR								Subgroup
van Beek[247]					✓									Narrative
van den Hoogen[248]			ROC	✓										Narrative
van der Wurff[249]				✓										Narrative
Varonen[251]	Test not described		ROC		✓	LR		Unweighted			✓		No	Subgroup
Vasbinder[252]								Weighting NS		✓			No	SROC regression
Verkooijen[253]	Fisher				✓									Narrative
Visser[254]	Statistical test NR; regression (predicted vs		Funnel plot – log DOR ROC		✓	Log DOR		Weighting NS				SROC parameters		SROC regression

continued

Study	Statistical tests			Method of study synthesis					If SROC give means of data presentation				For paired data: separate analysis?	Heterogeneity investigation: method
	Heterogeneity	Threshold effect	Graphical plot	Narrative	Pooled Se/Sp	Other pooled estimate	IPD	SROC	DOR	AUC	Q*	Other		
	observed value)													
Vroomen[255]					✓									Narrative
Watson[256]			ROC		✓	OR[g]							No	Subgroup
Wells[257]	Statistical test NR		ROC		✓	PV, LR		Method not described				Graph only		Subgroup
White[258]			Forest ROC		✓	PV, LR, accuracy		Weighting NS				Compared curves	No	Subgroup
Whited[259]				✓										Narrative
Whitsel[260]	χ^2		ROC		✓	DOR		Inverse variance weighted				*Se* at fixed *Sp*		SROC regression
			Forest ORs					Unweighted						
Wiese[261]	χ^2				✓								No	Subgroup
Wijnberger[262]		Spearman	ROC					Weighting NS				*Se* at 'almost perfect' *Sp* *Sp* at 95% *Se*		SROC regression; subgroup
Williams[263]	Statistical test NR		ROC			LR ES		Method not described				Graph only	No	

correl. coeff., correlation coefficient; DS&L, Dersimonian and Laird method; ES, effectiveness score; FPR, false positive rate; IPD, individual patient data; LR, likelihood ratio; PPV, positive predictive value; Se, sensitivity; Sp, specificity; TPR, true positive rate.

[a] Empirically weighted least squares.

[b] DOR-related – equivalent to log DOR.

[c] DOR estimated from logistic regression – akin to fixed effects pooling of DOR.

[d] Considered 'paired' data separately.

[e] Did Moses and colleagues' SROC first and decided SROC symmetric, so estimated SROC from pooled DOR instead.

[f] Youden index (δ) = sensitivity + specificity – 1.

[g] Likelihood of FN result on index vs reference test.

Appendix 6

Details of statistical investigations of sources of heterogeneity

Study	Quality-related items												Clinical/socio-demographic factors				Test-related			Summary information		
	Reference test used	Verification bias	Blinding	Disease progression	Sample description	Spectrum	Consecutive enrolment	Prospective/retrospective	Other item(s)	Quality score	Cohort/case-control	Sample size	Age	Sex	Prevalence/setting/spectrum	Other topic-specific	Test	Threshold	Publication year	No of variables	No of studies	Ratio variables: studies
Bachmann, 1998[64]																	✓			2	9	5
Badgett, 1997[50]	✓		✓				✓								✓			✓		2	34	17
Badgett, 1996[67]	✓		✓	✓		✓			✓						✓			✓		8	29	4
Bafounta, 2001[68]															✓		✓	✓		3	8	3
Balk, 2001[70]		✓	✓		✓										✓		✓			2	77	39
Bastian, 1998[74]																				1	5	5
Berger, 2000[79]	✓		✓			✓	✓							✓	✓					5	24	5
Berry, 1999[80]		✓	✓	✓	✓							✓					✓			5	7	1
Berry, 2002[82]		✓	✓		✓	✓	✓									✓	✓		✓	8	30	4
Blakeley, 1995[85]			✓		✓	✓	✓											✓		3	70	23
Buchanan, 2001[89]												✓			✓		✓			4	21	5
Buntinx, 1997[90]					✓			✓					✓		✓				✓	4	20	5
Carlson, 1994[93]															✓					1	34	34
Cher, 2001[94]		✓	✓				✓	✓		✓		✓			✓		✓			8	18	2
Chien, 1997[96]	✓		✓				✓		✓						✓					8	14	2
Choi, 2001[99]	✓							✓			✓							✓	✓	5	7	1
Conde-Agudelo, 1998[101]	✓		✓				✓						✓					✓		2	20	10
D'Arcy, 2000[52]	✓	✓											✓	✓	✓		✓			6	12	2
De Bernardinis, 1999[104]	✓									✓		✓						✓	✓	5	23	5
de Bruyn, 2001[105]	✓	✓	✓			✓		✓				✓				✓				4	12	3
de Vries, 1996[106]		✓	✓						✓						✓	✓	✓		✓	9	14	2
Deville, 2000[107]	✓	✓	✓		✓	✓		✓	✓	✓						✓		✓	✓	19	15	1
Devous, 1998[108]			✓										✓			✓	✓			6	30	5
Di Fabio, 1996[110]	✓														✓		✓			3	9	3
Dijkhuizen, 2000[213]	✓		✓				✓	✓							✓		✓	✓		7	39	6
Ebell, 2000[113]							✓									✓		✓		2	19	10
Eiberg, 2001[117]	✓																			1	28	28
Fahey, 1995[119]		✓	✓				✓								✓			✓	✓	5	62	12
Faron, 1998[120]															✓		✓			2	29	15

continued

Study	Quality-related items												Clinical/socio-demographic factors				Test-related			Summary information		
	Reference test used	Verification bias	Blinding	Disease progression	Sample description	Spectrum	Consecutive enrolment	Prospective/retrospective	Other item(s)	Quality score	Cohort/case-control	Sample size	Age	Sex	Prevalence/setting/ spectrum	Other topic-specific	Test	Threshold	Publication year	No of variables	No of studies	Ratio variables: studies
Fischer, 2001[123]	✓		✓		✓	✓		✓		✓							✓		✓	3	55	18
Fleischmann, 1998[124]		✓	✓										✓	✓	✓		✓	✓	✓	10	24	2
Gianrossi, 1990[128]		✓	✓	✓		✓							✓		✓		✓	✓	✓	12	13	1
Gould, 2001[132]			✓		✓		✓		✓	✓		✓					✓	✓	✓	12	40	3
Hallan, 1997[135]		✓													✓					1	22	22
Hobby, 2000[144]																	✓			1	16	16
Hoffman, 2000[145]		✓	✓			✓	✓			✓		✓			✓		✓	✓		7	17	2
Hofman, 2000[147]		✓						✓					✓		✓		✓			5	13	3
Hrung, 1999[149]		✓	✓		✓	✓		✓							✓		✓	✓	✓	4	16	4
Huicho, 2002[150]		✓	✓		✓	✓	✓	✓		✓			✓				✓	✓		4	48	12
Hurley, 2000[152]												✓			✓		✓			4	56	14
Ioannidis, 2001[154]		✓	✓		✓										✓					1	8	8
Ioannidis, 2001[155]		✓	✓		✓										✓		✓			2	17	9
Kearon, 1998[157]															✓					1	50	50
Kim, 2001[158]		✓	✓		✓					✓			✓	✓	✓		✓	✓	✓	10	82	8
Kinkel, 2000[159]								✓	✓						✓		✓		✓	12	46	4
Kinkel, 1999[160]		✓	✓		✓					✓					✓		✓		✓	6	47	8
Kittler, 2002[161]		✓	✓												✓		✓			7	27	4
Koelemay, 1996[164]		✓	✓		✓	✓	✓	✓		✓						✓		✓	✓	3	16	5
Koelemay, 2001[165]			✓		✓		✓	✓	✓						✓		✓			9	34	4
Koumans, 1998[53]		✓	✓		✓	✓	✓							✓	✓		✓			3	21	7
Kowalski, 2001[166]		✓	✓		✓		✓			✓					✓		✓		✓	4	16	4
Kwok, 1999[168]		✓	✓		✓					✓				✓	✓		✓	✓		9	21	2
Lacasse, 1999[169]		✓	✓					✓		✓					✓		✓			6	48	8
Law, 1998[172]		✓	✓		✓	✓	✓			✓			✓		✓		✓			4	45	11
Lederle, 1999[173]		✓	✓			✓	✓								✓		✓			1	15	15
Leitich, 1999[174]															✓		✓			6	28	5
Liedberg, 1996[176]										✓					✓					2	31	16
Loy, 1996[179]	✓	✓	✓		✓		✓					✓	✓		✓		✓			11	21	2

continued

Study	Quality-related items												Clinical/socio-demographic factors				Test-related			Summary information		
	Reference test used	Verification bias	Blinding	Disease progression	Sample description	Spectrum	Consecutive enrolment	Prospective/retrospective	Other item(s)	Quality score	Cohort/case-control	Sample size	Age	Sex	Prevalence/setting/spectrum	Other topic-specific	Test	Threshold	Publication year	No of variables	No of studies	Ratio variables: studies
Lysakowski, 2001[181]		✓	✓		✓		✓	✓							✓					1	26	26
Mackenzie, 1996[182]			✓					✓							✓					1	22	22
Markert, 1998[184]																		✓		1	7	7
McCrory, 1999[186]	✓	✓	✓		✓	✓	✓			✓					✓					10	109	11
Merritt, 1997[189]								✓									✓	✓		3	12	4
Mol, 1999[193]		✓													✓					2	25	13
Mol, 1998[44]							✓	✓			✓								✓	4	26	7
Mol, 1997[194]	✓		✓				✓										✓			4	23	6
Mol, 1998[195]		✓	✓				✓	✓							✓			✓		2	30	15
Mol, 1998[196]												✓					✓			2	24	12
Mushlin, 1998[201]																	✓			1	9	9
Nallamothu, 2001[203]		✓	✓									✓	✓	✓	✓				✓	6	14	2
Nanda, 2000[204]	✓	✓	✓			✓	✓								✓		✓			3	97	32
Oosterhuis, 2000[208]								✓		✓					✓		✓	✓		4	37	9
Orr, 1995[209]													✓	✓	✓				✓	5	17	3
Owens, 1996[55]	✓	✓	✓		✓	✓				✓		✓			✓			✓	✓	8	96	12
Owens, 1996[210]	✓	✓	✓		✓		✓			✓		✓	✓						✓	5	32	6
Peters, 1996[215]															✓					1	10	10
Rao, 1995[217]			✓			✓	✓	✓		✓					✓					2	15	8
Reed, 1996[221]	✓		✓						✓			✓				✓	✓			5	12	2
Revah, 1998[222]															✓					1	24	24
Safriel, 2002[224]												✓							✓	2	12	6
Scheid, 2001[225]																	✓			1	176	176
Scheidler, 1997[227]															✓					3	38	13
Scouller, 2000[229]	✓	✓	✓		✓	✓	✓						✓	✓	✓			✓		4	110	28
Siegman-Igra, 1997[231]																	✓			2	22	11
Smith-Bindman, 2001[233]											✓	✓						✓		3	56	19
Smith-Bindman, 1998[234]										✓					✓		✓	✓		5	35	7
Solomon, 2001[235]	✓		✓			✓	✓									✓	✓			2	23	12

continued

Study	Quality-related items											Clinical/socio-demographic factors				Test-related			Summary information		
	Reference test used	Verification bias	Blinding	Disease progression	Sample description	Spectrum	Consecutive enrolment	Prospective/retrospective	Other item(s)	Quality score	Cohort/case-control	Sample size	Age	Sex	Prevalence/setting/ spectrum	Other topic-specific	Test	Threshold	Publication year	No of variables	No of studies	Ratio variables: studies
Sonnad, 2001[236]		✓										✓					✓			5	27	5
Spencer-Green, 1997[237]	✓		✓		✓	✓				✓					✓		✓		✓	3	30	10
Stengel, 2001[238]	✓	✓	✓				✓									✓	✓			5	30	6
Swart, 1995[241]			✓	✓								✓			✓		✓			7	20	3
Tugwell, 1997[246]															✓			✓		1	8	8
Varonen, 2000[251]	✓		✓							✓										1	9	9
Vasbinder, 2001[252]			✓						✓			✓						✓	✓	5	55	11
Visser, 2000[254]			✓				✓	✓		✓			✓		✓		✓		✓	10	27	3
Watson, 2002[256]	✓		✓				✓	✓									✓			2	30	15
Wells, 1995[257]	✓		✓				✓	✓		✓							✓			2	16	8
White, 2000[258]			✓		✓							✓			✓				✓	5	38	8
Whitsel, 2000[260]			✓				✓					✓	✓	✓	✓			✓	✓	16	17	1
Wiese, 2000[261]		✓	✓			✓	✓	✓		✓					✓				✓	3	30	10
Wijnberger, 2001[262]		✓	✓				✓	✓									✓			1	6	6
Williams, 2002[263]		✓	✓		✓		✓										✓			1	14	14

Grey ticks indicate those items included in quality assessment but not as sources of heterogeneity.

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