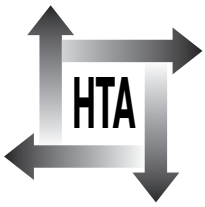
Health Technology Assessment 2007; Vol. 11: No. 50

Evaluation of diagnostic tests when there is no gold standard. A review of methods

AWS Rutjes, JB Reitsma, A Coomarasamy, KS Khan and PMM Bossuyt

December 2007

Health Technology Assessment
NHS R&D HTA Programme
www.hta.ac.uk



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.

Evaluation of diagnostic tests when there is no gold standard. A review of methods

AWS Rutjes,[1] JB Reitsma,[1] A Coomarasamy,[2] KS Khan[2*] and PMM Bossuyt[1]

[1] Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, The Netherlands
[2] Division of Reproductive and Child Health, University of Birmingham, UK

* Corresponding author

Declared competing interests of authors: none

Published December 2007

This report should be referenced as follows:

Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PMM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 2007;**11**(50).

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.

NIHR Health Technology Assessment Programme

The Health Technology Assessment (HTA) programme, now part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the costs, effectiveness and broader impact of health technologies 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 research findings from the HTA Programme directly influence decision-making bodies such as the National Institute for Health and Clinical Excellence (NICE) and the National Screening Committee (NSC). HTA findings also help to improve the quality of clinical practice in the NHS indirectly in that they form a key component of the 'National Knowledge Service'.

The HTA Programme is needs-led in that it fills gaps in the evidence needed by the NHS. There are three routes to the start of projects.

First is the commissioned route. Suggestions for research are actively sought from people working in the NHS, the public and consumer groups and professional bodies such as royal colleges and NHS trusts. These suggestions are carefully prioritised by panels of independent experts (including NHS service users). The HTA Programme then commissions the research by competitive tender.

Secondly, the HTA Programme provides grants for clinical trials for researchers who identify research questions. These are assessed for importance to patients and the NHS, and scientific rigour.

Thirdly, through its Technology Assessment Report (TAR) call-off contract, the HTA Programme commissions bespoke reports, principally for NICE, but also for other policy-makers. TARs bring together evidence on the value of specific technologies.

Some HTA research projects, including TARs, may take only months, others need several years. They can cost from as little as £40,000 to over £1 million, and may involve synthesising existing evidence, undertaking a trial, or other research collecting new data to answer a research problem.

The final reports from HTA projects are peer-reviewed by a number of independent expert referees before publication in the widely read monograph series *Health Technology Assessment*.

Criteria for inclusion in the HTA monograph series

Reports are published in the HTA monograph series if (1) they have resulted from work 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 National Coordinating Centre for Research Methodology (NCCRM), and was formerly transferred to the HTA Programme in April 2007 under the newly established NIHR Methodology Panel. The HTA Programme project number is 06/90/23. The contractual start date was in October 2005. The draft report began editorial review in March 2007 and was accepted for publication in April 2007. The commissioning brief was devised by the NCCRM who 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.

MR

Abstract

Evaluation of diagnostic tests when there is no gold standard. A review of methods

AWS Rutjes,[1] JB Reitsma,[1] A Coomarasamy,[2] KS Khan[2*] and PMM Bossuyt[1]

[1] Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, The Netherlands

[2] Division of Reproductive and Child Health, University of Birmingham, UK

* Corresponding author

Objective: To generate a classification of methods to evaluate medical tests when there is no gold standard.

Methods: Multiple search strategies were employed to obtain an overview of the different methods described in the literature, including searches of electronic databases, contacting experts for papers in personal archives, exploring databases from previous methodological projects and cross-checking of reference lists of useful papers already identified.

Results: All methods available were classified into four main groups. The first method group, impute or adjust for missing data on reference standard, needs careful attention to the pattern and fraction of missing values. The second group, correct imperfect reference standard, can be useful if there is reliable information about the degree of imperfection of the reference standard and about the correlation of the errors between the index test and the reference standard. The third group of methods, construct reference standard, have in common that they combine multiple test results to construct a reference standard outcome including deterministic predefined rules, consensus procedures and statistical modelling (latent class analysis). In the final group, validate index test results, the diagnostic test accuracy paradigm is abandoned and research examines, using a number of different methods, whether the results of an index test are meaningful in practice, for example by relating index test results to relevant other clinical characteristics and future clinical events.

Conclusions: The majority of methods try to impute, adjust or construct a reference standard in an effort to obtain the familiar diagnostic accuracy statistics, such as sensitivity and specificity. In situations that deviate only marginally from the classical diagnostic accuracy paradigm, these are valuable methods. However, in situations where an acceptable reference standard does not exist, applying the concept of clinical test validation can provide a significant methodological advance.
All methods summarised in this report need further development. Some methods, such as the construction of a reference standard using panel consensus methods and validation of tests outwith the accuracy paradigm, are particularly promising but are lacking in methodological research. These methods deserve particular attention in future research.

Contents

List of abbreviations

BCG	Bacillus Calmette–Guérin
CI	confidence interval
COPD	chronic obstructive pulmonary disease
CT	computed tomography
ED	emergency department
EIA	enzyme immunoassay
DOR	diagnostic odds ratio
FN	false negative result
FP	false positive result
LR	likelihood ratio
LTBI	latent tuberculosis infection
MAR	missing at random
MCAR	missing completely at random
MRI	magnetic resonance imaging
NMAR	not missing at random
NPV	negative predictive value
NT-proBNP	N-terminal pro-brain natriuretic peptide
PCR	polymerase chain reaction
PE	pulmonary embolism
PET	positron emission tomography
PPV	positive predictive value
RCT	randomised controlled trial
ROC	receiver operating characteristic
TN	true negative result
TP	true positive result
TST	tuberculin skin test

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

Executive summary

Background

The classical diagnostic accuracy paradigm is based on studies that compare the results of the test under evaluation (index test) with the results of the reference standard, the best available method to determine the presence or absence of the condition or disease of interest. Accuracy measures express how the results of the test under evaluation agree with the outcome of the reference standard. Determining accuracy is a key step in the health technology assessment of medical tests.

Researchers evaluating the diagnostic accuracy of a test often encounter situations where the reference standard is not available in all patients, where the reference standard is imperfect or where there is no accepted reference standard. We use the term 'no gold standard situations' to refer to all those situations. Several solutions have been proposed in these circumstances. Most articles dealing with imperfect or absent reference standards focus on one type of solution and discuss the strengths and limitations of that approach. Few authors have compared different approaches or provided guidelines on how to proceed when faced with an imperfect reference standard.

Objectives

We systematically searched the literature for methods that have been proposed and/or applied in situations without a 'gold' standard, that is, a reference standard that is without error. Our project had the following aims:

1. To generate an overview and classification of methods that have been proposed to evaluate medical tests when there is no gold standard.
2. To describe the main methods discussing rationale, assumptions, strengths and weaknesses.
3. To describe and explain examples from the literature that applied one or more of the methods in our overview.
4. To provide general guidance to researchers facing research situations where there is no gold standard.

Methods

We employed multiple search strategies to obtain an overview of the different methods described in the literature, including searches of electronic databases, contacting experts for papers in personal archives, exploring databases from previous methodological projects (STARD and QUADAS) and cross-checking of reference lists of useful papers already identified.

We developed a classification for the methods identified through our review taking into account the degree to which they represented a departure away from the classical diagnostic accuracy paradigm.

For each method in our overview, we prepared a structured summary based on all or the most informative papers describing its rationale, its strengths and weaknesses, its field of application, available software and illustrative examples of the method.

Based on the findings of our review, discussions about the pros and cons of different methods in various situations within the research team and input from expert peer reviewers, we constructed a flowchart providing general guidance to researchers faced with evaluation of tests without a gold standard.

Results

From 2200 references initially checked for their usefulness, we ultimately included 189 relevant articles that were subsequently used to classify and summarise all methods into four main groups, as follows.

Impute or adjust for missing data on reference standard

In this group of methods, there is an acceptable reference standard, but for various reasons the outcome of the reference standard is not obtained in all patients. Methods in this group either impute or adjust for this missing information in the subset of patients without reference standard outcome. Researchers should be careful with these

methods if (1) the pattern of missing values is not determined by the study design, but is influenced by the choice of patients and physicians, or (2) the fraction of patients verified with the reference standard is small within results of the index tests.

Correct imperfect reference standard

In this group, there is a preferred reference standard, but this standard is known to be imperfect. Solutions from this group either adjust estimates of accuracy or perform sensitivity analysis to examine the impact of this imperfect reference standard. The adjustment is based on external data (previous research) about the degree of imperfection. Correction methods can be useful if there is reliable information about the degree of imperfection of the reference standard and about the correlation of the errors between the index test and the reference standard.

Construct reference standard

These methods have in common that they combine multiple test results to construct a reference standard outcome. Groups of patients receive either different tests (differential verification and discrepant analysis) or the same set of tests, after which these results are combined by: (1) deterministic predefined rule (composite reference standard); (2) consensus procedure among experts (panel diagnosis); (3) a statistical model based on actual data (latent class analysis). The prespecified rule for target condition makes the composite reference standard method transparent and easy to use, but misclassification of patients is likely to remain. Discrepant analysis should not be considered in general, as the method is likely to produce biased results. The drawback of latent class models is that the target condition is not defined in a clinical way, so there can be lack of clarity about what the results stand for in practice. Panel diagnosis also combines multiple pieces of information, but experts may combine these items in a manner that more closely reflects their own personal concept of the target condition.

Validate index test results

The diagnostic test accuracy paradigm is abandoned in this group and index test results are related to relevant other clinical characteristics. An important category is relating index test results with future clinical events, such as the number of events in those tested negative for the index test results. Test results can also be used in a randomised study to see whether the test can predict who will benefit more from one intervention than the other. Because the classical accuracy paradigm is not employed, measures other than accuracy measures are calculated, including event rates, relative risks and other correlation statistics.

Conclusions

The majority of methods try to impute, adjust or construct a reference standard in an effort to obtain the familiar diagnostic accuracy statistics such as pairs of sensitivity and specificity or likelihood ratios. In situations that deviate only marginally from the classical diagnostic accuracy paradigm, for example where there are few missing values on an otherwise acceptable reference standard or where the magnitude and type of imperfection in a reference standard is well documented, these are valuable methods. However, in situations where an acceptable reference standard does not exist, holding on to the accuracy paradigm is less fruitful. In these situations, applying the concept of clinical test validation can provide a significant methodological advance. Validating a test means that scientists and practitioners examine, using a number of different methods, whether the results of an index test are meaningful in practice. Validation will always be a gradual process. It will involve the scientific and clinical community defining a threshold, a point in the validation process, whereby the information gathered would be considered sufficient to allow clinical use of the test with confidence.

Recommendations for further research

All methods summarised in this report need further development. Some methods, such as the construction of a reference standard using panel consensus methods and validation of tests outwith the accuracy paradigm, are particularly promising but are lacking in methodological research. These methods deserve particular attention in future research.

Chapter 1
Background

"Accuracy is telling the truth … Precision is telling the same story over and over again."

Yiding Wang

Introduction and scope of the report

As with all other elements of healthcare, medical tests should be thoroughly evaluated in high-quality studies. Biased results from poorly designed, conducted or analysed studies may trigger premature dissemination and implementation of a medical test and mislead physicians to incorrect decisions regarding the care for an individual patient.[1–3] Avoidance of these perils requires a proper evaluation of medical tests.[4]

Several authors have proposed a staged model for the evaluation of medical tests.[5–7] Technical evaluations dominate in the early phases, in which the reproducibility under different conditions of biochemical tests and the intra- and inter-observer variation of tests are evaluated. A key phase in the clinical evaluation of a test is determining its diagnostic accuracy: the ability to discriminate between patients who have the condition of interest (target condition) and those who have not.[8] The target condition can refer to a disease, syndrome or any other identifiable condition that may prompt clinical actions such as further diagnostic testing, or the initiation, modification or termination of treatment.

By themselves, accuracy studies cannot always answer the question whether a medical test is useful or not. More informative accuracy studies can be designed by taking into account the likely future role of the test under evaluation. Three possible roles are replacement, addition or triage.[9] In comparative accuracy studies, the accuracy of the test under evaluation is compared against that of existing diagnostic pathways, leading to more informative and possibly more efficient diagnostic accuracy studies.[9]

The clinical value of a test will ultimately depend on whether it is able to improve patient outcome. In most cases this will be by guiding subsequent decision-making. Accuracy studies may not be sufficient to evaluate the clinical value of a test, especially if the new test is more sensitive than the existing test(s).[10] The reason is that the results from current intervention studies may not apply to those additional cases detected. Results from randomised studies assessing response to therapy in these additional cases are then required. Later stage evaluation studies may focus on determining the societal costs and benefits of a test.

The focus of this report is on problems related to diagnostic accuracy studies. The key challenge in diagnostic accuracy is to determine in all patients whether the target condition is present or absent. The reference standard should provide this classification. As such, the reference standard plays a crucial role in accuracy studies.

Problems with the reference standard abound in diagnostic accuracy studies. (http://www.fda.gov/cdrh/osb/guidance/1428.pdf). The outcome of the reference standard may not be available in all patients, it may be unreliable, it may be inaccurate or there could be no acceptable reference standard at all. As in any other form of epidemiological research, outcome data that are missing or misclassified pose a great threat to the validity of such studies, and diagnostic accuracy studies are no exception.[11–15] We will use the term 'no gold standard situations' to refer loosely to all these situations where the outcome of the reference standard is missing, the reference standard is imperfect or there is no acceptable reference standard.

This report gives an overview of solutions that have been proposed to overcome no gold standard situations in diagnostic accuracy research. Because the scope of this report is limited to problems related to diagnostic accuracy studies, we will not discuss the broader issue of determining the most appropriate type of evaluation given a specific diagnostic research question.

Before explaining our methods (Chapter 3) and reporting our results (Chapter 4), in the next section we first explain the key features of diagnostic accuracy studies and in the subsequent section discuss the various mechanisms that can

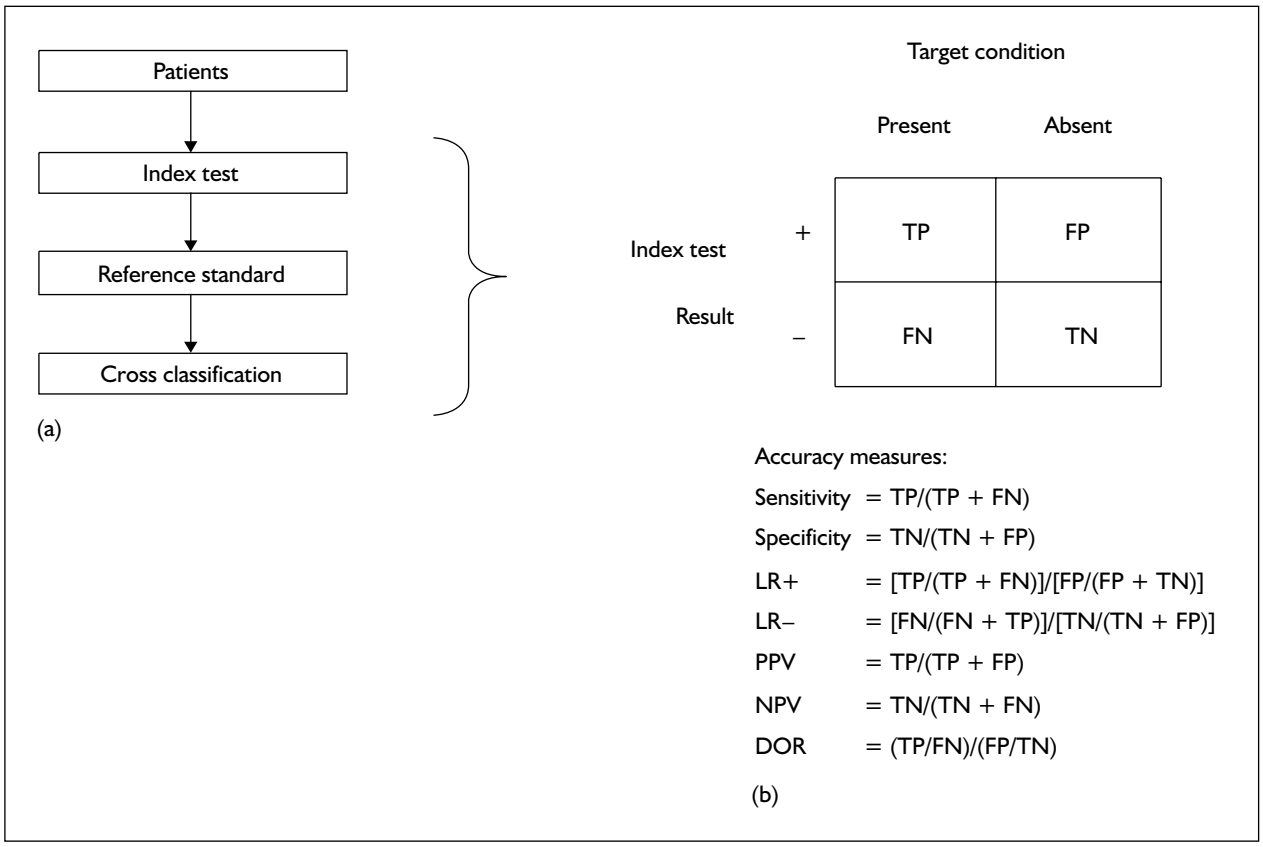


FIGURE 1 *(a) Classical design of a diagnostic accuracy study and (b) results of an accuracy study in the case of a dichotomous index test result. TP, true positive result; FP, false positive result; FN, false negative result; TN, true negative result; LR, likelihood ratio; PPV, positive predictive value; NPV, negative predictive value; DOR, diagnostic odds ratio.*

lead to no gold standard situations. In Chapter 5 we provide guidance on selecting the most appropriate method for a given situation by addressing some key questions. In addition, we sketch possible research alternatives in situations where the diagnostic accuracy paradigm is unlikely to be useful.

Key concepts in diagnostic accuracy studies

Diagnostic accuracy studies aim to measure the amount of agreement between index test results and the outcome of the reference standard. The classical outline of an accuracy study is given in *Figure 1(a)*. The starting point is a consecutive series of individuals in whom the target condition is suspected. The index test is performed first in all subjects, and subsequently the presence or absence of the target condition is determined by the outcome of the reference standard. In the case of a dichotomous index test result, the results of an accuracy study can be summarised in a 2-by-2 table, as shown in *Figure 1(b)*. Several measures of accuracy can be calculated from this table, as also shown in *Figure 1(b)*.

The term accuracy has been borrowed from measurement theory, where it is defined as the closeness of agreement between an analytical measurement and its actual (true) value.[16] The first publication mentioning accuracy and the associated statistics sensitivity and specificity to express the performance of a medical test was by Yerushalmy,[17] followed by the landmark publication of Ledley and Lusted.[18] In these publications, test results in patients known to have the disease of interest were compared with test results in subjects not having the disease. Since then, various other measures of accuracy have been introduced, including likelihood ratios, predictive values and the diagnostic odds ratio.[4] All of these accuracy measures have in common that they need a classification of patients in those with and those without the condition of interest. In other words, index test results are verified by comparing them with the outcome of the reference standard. Based on this concept, we can formulate the properties of the ideal reference

standard and verification procedure. The ideal verification protocol would fulfil the following criteria:

1. The reference standard provides error-free classification.
2. All index test results are verified by the same reference standard.
3. The index test and reference standard are performed at the same time, or within an interval that is short enough to eliminate changes in target condition status.

Empirical studies have shown that estimates of diagnostic accuracy are directly influenced by the quality of the verification procedure.[1,3,19]

Problems in verification

In practice, researchers may encounter situations where the ideal verification procedure cannot be achieved. Based on the first two criteria for an ideal verification procedure, we will briefly discuss the key mechanisms why verification procedures can produce errors.

Classification errors by the reference standard

Within the accuracy concept, the reference standard is judged by its performance in producing error-free classification with respect to the presence or absence of the target condition. Even if a reference standard has no analytical error but the target condition does not produce the biochemical changes of interest, we consider it a 'failure' of the reference standard. Other examples of imperfections of the reference standard are tumours that are missed because they are below the level of detection in case of a radiological reference standard, or the presence of an alternative condition that is misclassified as the target condition because it produces similar changes in a biomarker as that of the target condition.

One inherent difficulty in accuracy studies is that we use the dichotomy of target condition present or absent, whereas in reality the target condition varies from very early and minor changes to severe and advanced stages of the disease, thus covering a wide spectrum of disease. For some conditions, defining the lower end of disease is difficult. This can be illustrated with the example of appendicitis. In the majority of accuracy studies evaluating non-invasive imaging techniques for the diagnosis of appendicitis, histological examination of the removed appendix is used as reference standard. This requires defining a threshold for the type and amount of inflammatory changes above which we will classify a patient as having appendicitis. Different reference standards may apply a different threshold before classifying patients as having the target condition. In the example of appendicitis, clinical follow-up and observing whether the symptoms of the patients improve or deteriorate will probably mean a higher threshold for disease, resulting in some patients being classified differently between these two reference standards; for example, some patients with inflammatory changes at histological examination would have recovered in a natural way without intervention. This requires proper thinking by the researcher of what is the right definition of the target condition and then choosing the most appropriate reference standard in the light of this target condition.

Additional sources of misclassification are failures in the reference standard protocol and interpretation errors by observers. These errors in misclassification could be preventable by stricter adherence to protocol or better training of observers. Examples include failure of detecting cancer cells after fine-needle aspiration because the biopsy was performed outside the tumour mass or overlooking a small pulmonary embolism in spiral computed tomography (CT) images.

Furthermore, for several target conditions there is no reference standard based on histological or biochemical changes. In some of those cases, the condition is defined by a combination of symptoms and signs. Migraine is an example. The presence or absence of this and similar conditions has been based on criteria developed by individual researchers or on criteria established during a consensus meeting. Such classifications can vary over time or across countries, and cannot be error-free.

Given the many potential factors that can lead to errors in the classification of the target condition by a reference standard, a perfect reference standard (e.g. providing error-free classification) is unlikely to exist in practice. The shift in terminology from 'gold standard', suggesting a standard without error, to the more neutral term 'reference standard', indicating the best available method, has been initiated by these observations. It means that researchers should always discuss the quality of the reference standard and the potential consequences of misclassification by the reference standard.

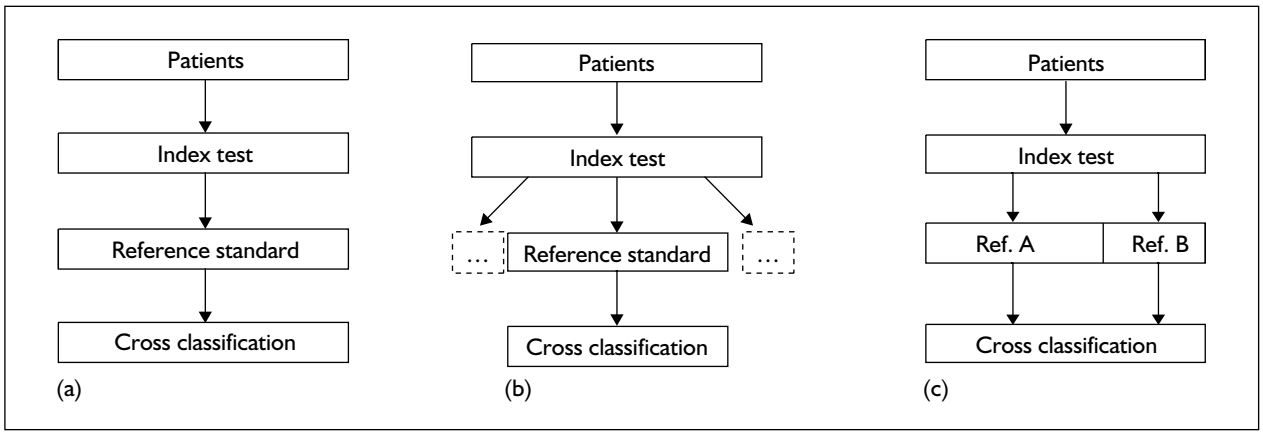


FIGURE 2 *(a) Diagnostic accuracy study with complete verification by the same reference standard (classic design), (b) study with partial verification and (c) study with differential verification.*

Whatever the cause of errors in classification of target condition status by the reference standard, it will directly lead to changes in the 2-by-2 classification table (*Figure 1b*). Within the classic accuracy concept, any disagreement between the reference standard and the index test will be labelled as a 'false' result for the index test. The net effect of the misclassification by the reference standard can either be an upward or downward bias in estimates of diagnostic accuracy. The direction depends on whether errors by the index test and imperfect reference standard are correlated. If errors are positively correlated, it will erroneously increase agreement in the 2-by-2 tables and estimates of accuracy will be inflated.[11–15] The magnitude of the biasing effect depends on the frequency of errors by the imperfect reference standard and the degree of correlation in errors between index test and reference standard.

Partial and differential verification

Even if a near perfect reference standard exists, it may be impossible, unethical or too costly to apply this standard in all patients. In the case of target conditions that can produce multiple lesions that need histological verification, it is often impossible to verify (the countless) negative index test results. An example is 18-fluorodeoxyglucose-positron emission tomography (PET) scanning to detect possible distant metastases before planning major curative surgery in patients with carcinoma of the oesophagus: only PET hot spots can be verified histologically.

Ethical reasons play a role in the choice of reference standard in pulmonary embolism. Angiography is still considered the best available method for the detection of pulmonary embolisms, but because of the frequency of serious complications it is now considered unethical to perform this reference standard in low-risk patients, for instance in patients with low clinical probability and negative D-dimer result. Other reasons for missing data on the outcome of the reference standard are situations where the reference standard is temporarily unavailable or when patients and doctors decide to refrain from verification.

One solution is to leave the unverified patients out of the 2-by-2 table, which is referred to as partial verification (*Figure 2b*). Omitting unverified patients from the 2-by-2 table may generate a bias. The direction and magnitude will depend on (1) the fraction of patients that are unverified; (2) the ratio between the number of patients with positive and negative index test results that remain unverified; and (3) whether the reason for not verifying is related to the presence or absence of the target condition.[20]

Because omitting patients from the 2-by-2 classification table can lead to bias, researchers have strived to obtain complete verification by applying an alternative reference standard in the initially unverified patients. The use of different reference standards between patients is known as differential verification (*Figure 2c*). An example of differential verification that is fairly common is where follow-up is used as the alternative reference standard in patients not verified by the preferred reference standard. In these studies, clinical follow-up is used as a proxy to obtain the information of true status at the moment of the index test; the term delayed-type cross-sectional accuracy study has been introduced for this design.[8] For all studies with differential verification where the alternative reference standard provides imperfect classification, it may affect the 2-by-2

table and generate bias along the same lines as discussed earlier in the previous section on errors in classification by the reference standard. Empirical studies have shown that studies with differential verification produce higher estimates of diagnostic accuracy than studies with complete verification by the preferred reference standard.[1,3]

Given these diverging reasons for imperfections in the verification procedure, it is not surprising that different solutions have been suggested to remedy the effects of verification procedures that are not based on using a single gold standard in all patients. We have systematically searched the literature to identify and summarise the solutions that have been proposed. We have developed a classification of these methods based on their figurative distance, that is, the extent to which they depart from the classical diagnostic accuracy paradigm described in *Figure 1*.

Chapter 2

Aims of the project

Several methods have been proposed to deal with situations where the reference standard is partially unavailable or imperfect or where there is no accepted reference standard. These methods vary from relatively simple correction methods based on the expected degree of imperfection of the reference standard, through more complex statistical models that construct a pseudo-reference standard, to methods that validate index test results by examining their association with other relevant clinical characteristics. Most articles dealing with imperfect or absent reference standards focus on one type of solution and discuss strengths and limitations of that particular approach.[21–23] Few authors have compared different approaches or have provided guidelines on how to proceed when faced with an imperfect reference standard[24–26] (http://www.fda.gov/cdrh/osb/guidance/1428.pdf).

Aims

In this project, we systematically searched the literature for methods to be used in situations without a 'gold' standard. Our project had the following aims:

1. To generate an overview and classification of methods that have been proposed to evaluate medical tests when there is no gold standard.
2. To describe the main methods discussing rationale, assumptions, strengths and weaknesses.
3. To describe and explain examples from the literature that applied one or more of the methods in our overview.
4. To provide general guidance to researchers facing research situations where there is no gold standard.

Outline of the report

Chapter 1 provides background information about the key concepts in diagnostic accuracy studies, in particular the role of the reference standard and the problems that can be encountered in the verification of index tests results. The methods chapter (Chapter 3) describes the strategies that we employed to identify and assess methodological papers on methods relevant for this project. The results chapter (Chapter 4) has the following structure: the first section presents the results of the literature search; an overview and classification of the different methods that we encountered is given in the second section; and a description of each of the methods discussing strengths and limitations is given in the third section. In Chapter 5 we provide general guidance to researchers when faced with research situations where there is no gold standard. In addition, we discuss some alternative options when repairing or hanging on to the accuracy paradigm is not helpful.

Chapter 3
Methods

In our systematic review, we modified the recommendations set out by the Cochrane Collaboration and the NHS Centre for Reviews and Dissemination[27] to make them more applicable for a review of methodological papers.

Our approach consisted of the following steps:

1. literature search and inclusion of papers
2. assessment of individual studies
3. overall classification of methods
4. structured summary of individual methods
5. expert review on individual methods
6. development of guidance statements
7. peer review of total report.

Literature search and inclusion of papers

Searching for methodological papers in electronic databases is difficult because of inconsistent indexing and the absence of a specific keyword for the relevant publication types. Several methods for conducting diagnostic research when there is no gold standard have been developed in other areas of epidemiological, biomedical or even non-medical research. Conducting a broad literature search to capture every single potentially relevant paper would be inappropriate, especially since precise estimation is not the objective of a review of methodological papers. Once a set of comprehensive papers has been obtained about a methodological issue, there is no additional value in reviewing additional papers explaining the same concept. This is known as theoretical saturation,[28] a principle that guided our search and selection.

Based on these considerations, we relied on multiple strategies to obtain an overview of the different methods described in the literature and to maximise the likelihood of identifying those papers that provide the most thorough and complete description:

- Restricted electronic searches in the following databases: MEDLINE (OVID and PubMed), EMBASE (OVID), MEDION, a database of diagnostic test reviews (www.mediondatabase.nl), and the Cochrane Library (DARE, CENTRAL, CMR, NHS). The exact terms of the search strategy are given in Appendix 1.
- Searching databases that have been established in earlier methodological projects, including STARD and QUADAS.
- Searching personal archives.
- Contacting other experts in the field of diagnostic research to establish whether they had additional relevant papers, especially aimed at retrieving articles within the grey literature.
- To locate additional information on specific methods, we checked reference lists, used the citation tracking option of SCISEARCH and applied the 'related articles' function of PubMed.

Inclusion of papers

Publications in English, German, Dutch, Italian, and French were included. One researcher (AWSR) reviewed the identified studies for inclusion in the review and this process was checked by another reviewer (JBR). The only reason for exclusion was if the article did not address a method that could be used in a no gold standard situation.

Assessment of individual studies

We extracted a limited set of standard items from each included paper. These included the strategy that identified the paper, the type of journal, the type of article and the type of method proposed. Other items were the rationale and structure of the article, a description of its usefulness and remarks about the overlap in relation to other papers (all free text fields). The information extracted in this way was used to organise papers within each method. The main function of this step was to categorise papers in order to facilitate the writing of a structured summary of each method.

Overall classification of methods

We classified all methods into groups that addressed diagnostic research in situations where there is no gold standard in an analogous way. The purpose of this classification was to give an

overview of methods and to serve as a starting point for formulating guidance. The key element in the classification was the underlying mechanism leading to the absence of perfect reference standard information (see also Chapter 1).

Structured summary of individual methods

For each method in our overview, we made a structured summary based on all or the most informative papers describing that method. Each summary starts with a description of the method and its rationale, avoiding technical language where possible. In subsequent sections, the strengths and weaknesses of the methods are described; the field of application of the method in terms of the circumstances in which the method should or should not be used; available software; and an illustrative example of the method. One member of the research team (AWSR or AC) wrote the first draft of a summary, which was then reviewed and modified by another team member (JBR, PMMB, AC, KSK). These two reviewers continued exchanging drafts, held face-to-face meetings or had teleconference meetings until they agreed that the version was ready for expert review.

Expert review on individual methods

We assembled a list of potential reviewers outside our research team to review each method. These experts were selected based on their knowledge and/or expertise in relation to a specific method. This group included epidemiologists, statisticians and clinicians (see Appendix 2). Each expert was contacted by electronic mail or by telephone and was asked to comment on at least one method. These experts received a brief description of the project, its objectives and the summary of a specific method. They were asked to review this first version, paying particular attention to the following issues:

- Is the method accurately described?
- Are the main characteristics of the method described?
- Are key strength and weaknesses mentioned?
- Are key references missing?
- Are tables and/or figures understandable?
- Do you have any other suggestions how the summary can be improved?

Based on the comments from the expert, a final version of each method was prepared. Because of time constraints, the revised version was not sent back to the expert.

Development of research guidance

The research team held face-to-face meetings to discuss the pros and cons of different methods in different situations. Prior to this meeting, we had contacted and interviewed a few general experts about their preferred solutions when faced with research situations where there is no gold standard. Based on the discussion, a first version was drafted and circulated among the research members. Further comments were incorporated to produce a final version.

Peer review of report

The full report was reviewed by general experts identified by the research team and peer reviewers assigned by the NHS Research Methodology programme.

Chapter 4

Results

Search results and selection of studies

The references retrieved by the electronic searches were assembled in a single database. After duplicate references had been deleted, the abstracts of 2265 references were assessed for eligibility. Of those, 134 were labelled potentially relevant. Twenty-three references were excluded because they could not be retrieved or inspection of the full article revealed that the topic was not test evaluation. Contact with experts in the field resulted in an additional seven references to books and 50 reference to articles. Reference checking at various stages of preparing the report yielded a total of 11 additional relevant articles.

As the number of references was limited for imputation methods and the panel consensus method, we conducted additional electronic searches. Using the 'related articles' approach in PubMed we identified 10 additional references relevant for imputation. For the panel consensus method, an additional search in PubMed was performed with the search term Delphi[textword], resulting in no relevant references. The final number of relevant articles included was 189. A flowchart of the retrieval and inclusion of studies is given in *Figure 3*.

Classification of methods

Many methods have been described for no gold standard situations. An exact number is difficult to provide because many methods can be viewed as variations or extensions of a single underlying approach. For this report, we classified the methods into four main groups (*Table 1*). These four groups differ in their distance or extent of

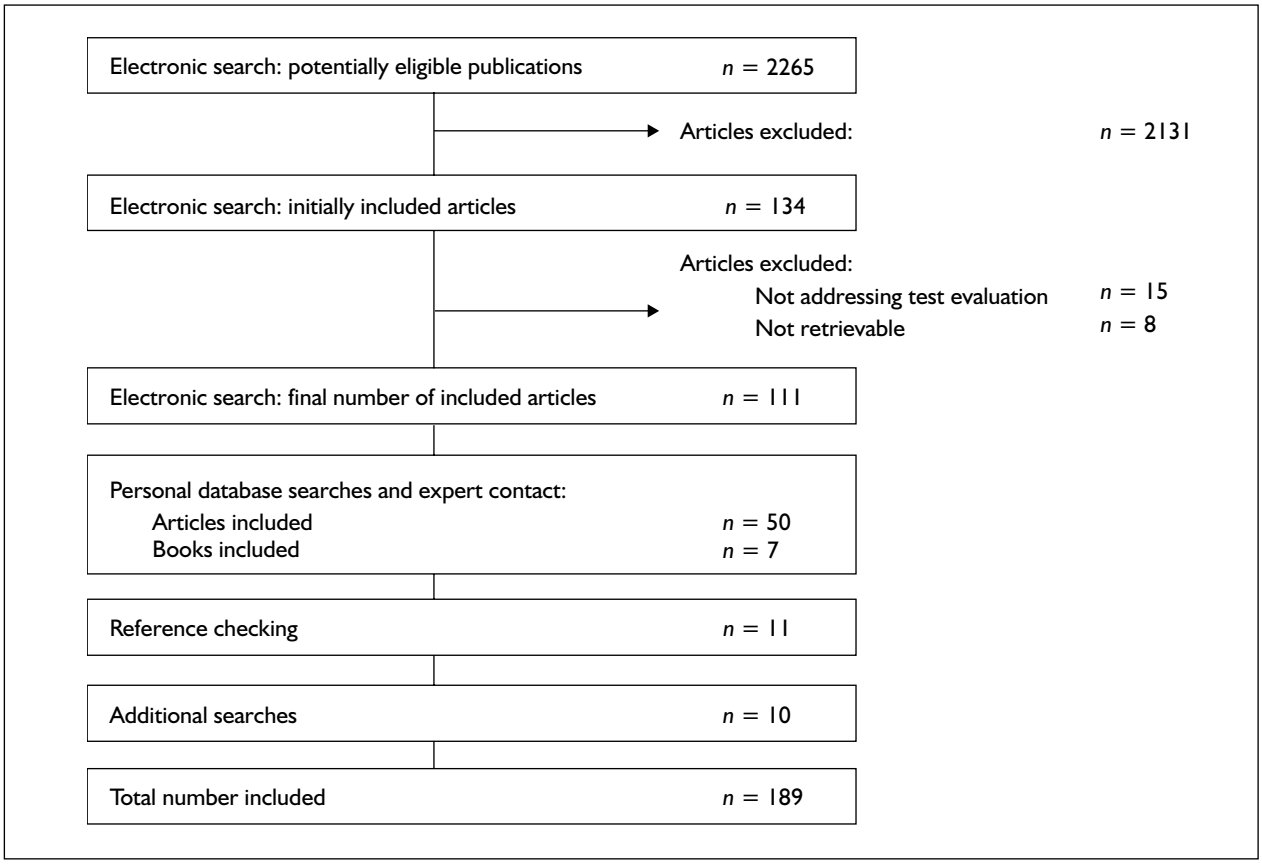


FIGURE 3 *Flowchart of the selection process of articles*

TABLE 1 *Classification of methods for diagnostic research where there is no gold standard*

Main classification *Subdivision*	**Main characteristic**	**Section in Chapter 4**[a]
A. Impute or adjust for missing data on reference standard	Impute the outcome of the reference standard in those patients who did not receive verification by the reference standard or adjust estimates of accuracy based on complete cases	A
B. Correct imperfect reference standard	Correct estimates of accuracy or perform sensitivity analysis to examine the impact of using imperfect reference standard based on external data about the degree of imperfection	B
C. Construct reference standard *Differential verification* *Discrepant analysis* *Composite reference standard* *Panel or consensus diagnosis* *Latent class analysis*	Information from different tests is combined to construct the reference standard outcome. Groups of patients receive either different tests (differential verification and discrepant analysis) or the same set of tests after which these results are combined by: (a) deterministic predefined rule (composite reference standard) (b) consensus procedure among experts (panel diagnosis) (c) a statistical model based on actual data (latent class analysis)	C D E F G H
D. Validate index test results *Examine patient outcomes*	Explore meaningful relations between index test results and other relevant clinical characteristics. An important way to validate is to use dedicated follow-up to capture clinical events of interest in relation to index test results, including randomised diagnostic studies	I

[a] A, 'Impute or adjust for missing data on reference standard' (p. 13); B, 'Correct imperfect reference standard' (p. 16); C, 'Construct reference standard' (p. 18); D, 'Differential verification' (p. 18); E, 'Discrepant analysis' (p. 19); F, 'Composite reference standard' (p. 21); G 'Panel or consensus diagnosis' (p. 24); H, 'Latent class analysis' (p. 26); I, Validate index test results' (p. 29).

departure from the classical diagnostic accuracy paradigm including a gold standard, as explained in Chapter 1. Our classification is neither exhaustive nor mutually exclusive. Researchers can also use a combination of methods within a single study.

In the first group of methods, **Group A: Impute or adjust for missing data on reference standard**, there is a reference standard providing adequate classification, but for various reasons the outcome of the reference standard is not obtained in all patients (see Chapter 1 for an overview of reasons). The methods in this group impute or adjust for this missing information in the subset of patients without reference standard outcome. Methods within this group differ in the way in which they impute or adjust for this missingness.

In the second group, **Group B: Correct for imperfections in reference standard**, there is a preferred reference standard, but this standard is known to be imperfect. Solutions from this group either adjust estimates of accuracy or perform sensitivity analysis to examine the impact of using this imperfect reference standard. The adjustment is based on external data (previous research) about the degree of imperfection.

Methods in the third group, **Group C: Construct a reference standard**, have in common that they combine multiple pieces of information (test results) to construct a reference standard outcome. The methods differ as to whether they use a predefined deterministic rule to classify patients as having the target condition (composite reference standard), whether only discordant results are retested with a second reference standard (discrepant analysis) or whether the different tests are combined through a statistical model (latent class analysis). Also, a panel of experts can be used to determine the presence or absence of the target condition in each patient.

The diagnostic test accuracy paradigm is abandoned in the fourth group, **Group D: Validate index test results**. In these studies, index test results are related to other relevant clinical characteristics. An important category is relating index test results with future clinical events, such as the number of events in those tested negative for the index test or a randomised comparison between testing and non-testing. Any relevant clinical information could be used to validate index test results, including cross-sectional or historical data. Because this group departs completely from the classical accuracy paradigm,

measures other than accuracy measures are calculated, including event rates, relative risks and correlation statistics.

Impute or adjust for missing data on reference standard

In some studies, an accepted reference standard providing adequate classification is available, but for a variety of reasons not all patients receive this standard, leading to missing data on whether the target condition is present or absent. The general problem of missing data is well established in epidemiology and biostatistics, and many statistical methods have been developed to deal with missing data.[29] Most literature focuses on missing values on predictors, but a fair amount of literature is available on missing data of outcome variables, including papers dealing with diagnostic research.[22,23,30–41] Missing data on the reference standard have been labelled partial or incomplete verification in the diagnostic literature, and the bias associated with it is known as partial verification bias or sequential ordering bias.[19,20]

Imputation methods use a mathematical function to fill in each missing value, whereas correction methods use a mathematical function to correct the indexes of accuracy directly. Three main patterns of partial verification or missingness have been described: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR).[29]

In MCAR, missing values are unrelated to the status of the target condition, index test results and other patient characteristics. Hence the occurrence is a true random process and therefore the frequency of missing values is similar among patients with positive or negative index test results. This situation can occur due to unavailability of the reference standard because of technical failures.

More often, missing values for the reference standard do not occur completely at random, but are related to the results of the index test (more frequent in patients with negative index test result) and to other patient characteristics. If the mechanisms that have led to the missing values on the reference standard are known and observed, then the technical term 'missing at random' (MAR) is used. This term is used because within strata of the mechanisms leading to missing values, the pattern is random. This also forms the basis for predicting and subsequently imputing missing values.

If the pattern of missing values is related to the target condition, but through mechanisms that we have not observed, it is called NMAR. This situation can occur when incomplete verification is not design based, but determined by the choice of patients and physicians. In these situations, the mechanisms leading to missing values are difficult to determine, especially when additional patient characteristics such as symptoms, signs or previous test results have not been recorded.

Imputation methods

Imputation methods comprise two phases, an imputation phase where each missing value is replaced, and an analysis phase, where estimates of sensitivity and specificity are computed based on the now complete dataset. Many variants of imputation are possible, ranging from single imputations of missing values to multiple imputations.[42–46] Instead of filling in a single value for each missing value, multiple imputation procedure replaces each missing value with a set of plausible values that represent the uncertainty about the correct value to impute. These multiple imputed data sets are then analysed (one by one) by standard procedures for complete data sets. In a next step, the results from these analyses are combined to produce estimates and confidence intervals (CIs) that properly reflect the uncertainty due to missing values.

The choice of imputation method is largely based on the pattern of missingness. Basically, the associations between patient characteristics and the outcome of the reference standard are evaluated through a statistical model, like logistic regression, subsequently to impute missing values. In the NMAR setting, it is very rare to identify an appropriate model for the missingness mechanism. Consequently, the validity of the model is uncertain. For further information on dealing with missingness and imputations, readers are referred to Little and Rubin.[29]

Correction methods for missing data on reference standard outcome

In correction methods, no effort is made to impute missing values. The notion is that estimates of diagnostic test accuracy are likely to be biased if only patients who received verification with the reference standard (complete cases) are analysed. A mathematical correction of these

estimates and their CIs is warranted based on the mechanism of missingness.

The first publications about correction for partial verification were based on the MCAR assumption.[39] If missings occur completely at random and the fractions of persons missing in each of the cells of the two-by-two table are likely to be comparable, estimates (but not CIs) of accuracy can be easily recalculated without any statistical model. As this pattern is not likely to occur often, mathematical correction methods have been developed that can deal with MAR. In MAR, estimates of accuracy can be corrected if verification is random within specific patient profiles, based on patient characteristics or (index) test results, and the number of patients not verified within each stratum is known. Here, several methods to correct estimates of sensitivity and specificity based on this assumption have been described.[13,22,23,32,47,48] The same is true for the correction of receiver operating characteristic (ROC) curves and associated indexes.[38,47,49]

Strengths and weaknesses

Both methods aim to alleviate the bias that is potentially introduced by analysing only cases with complete data on the reference standard. The strengths of each method are determined by how accurately the mechanism behind missing values is known. If the mechanisms that lead to missing values on the reference standard are (partially) unknown, the correction or imputation methods are prone to bias. Moreover, correction and imputation methods are based on statistical modelling of the data, which requires sufficiently large sample sizes.[50] The potential of bias can be large when correction methods are used in studies with relative small sample sizes, or if the fraction of patients receiving the reference standard is small. Imputation methods, especially multiple imputation methods, seem to be more robust in these situations than correction methods.[23]

Application

These methods can be used when a preferred reference standard is available, but has not been applied in all patients enrolled in the study. Ideally, partial or incomplete verification should be planned by design, so that the pattern of missingness is known. Researchers should be very careful with these methods if missingness is uncontrolled and open to influence by patient and practitioner choice, if the sample size is small or if the fraction of verified patients is small within test categories.

Software

Software for dealing with missing data is now embedded in the main statistical software packages.

Clinical example

Harel and Zhou used two real data examples to illustrate the results of multiple imputation and correction methods for missing data on the outcome of the reference standard.[23] We present a brief summary of their results.

The first example addresses hepatic scintigraphy, an imaging scan procedure, for the detection of liver cancer.[51] Out of 650 patients, 344 were referred to liver pathology, which was considered the reference standard (*Table 2*). The MCAR assumption does not hold here, as 39% of the index test positives and 63% of the index test negatives are not verified. Either MAR or NMAR is true, and Harel and Zhou used the MAR assumption that non-verification occurred randomly within index test positives and negatives, respectively. *Table 3* presents summary estimates of sensitivity, specificity and CIs as computed by eight different methods. The first method concerns a 'complete case' analysis, where the 306 patients without liver pathology are ignored and estimates are based on the remaining 344 patients only. This method gives a higher estimate of sensitivity and lower estimate of specificity in comparison with all other methods. The two correction methods, as proposed by Begg and Greenes,[22,48] produce comparable estimates, with lower sensitivities and higher specificities compared with the complete case analysis. In comparison with the remaining five multiple imputation methods, however, the sensitivities are lower and the specificities are higher.

In the second example, Harel and Zhou used data of a study evaluating diaphanography as a test for detecting breast cancer.[52] Of the 900 patients enrolled, 812 were not verified. The pattern of missingness was either MAR or NMAR, as 55% of the diaphanography positives and only 6% of the diaphonography negatives were verified (*Table 4*). Again, Harel and Zhou chose to use the MAR assumption that is related to index test results only. *Table 5* shows the estimates computed by the eight methods. Here the differences between the complete case analysis, the correction and imputation methods are more prominent, while the estimates of multiple imputations are fairly close to each other. In the light of the sample size and their previous simulation work, Harel and Zhou concluded that the estimates of multiple imputations are more representative of the data.

TABLE 2 *Hepatic scintigraphy data: outcome of reference standard in those verified and fraction of unverified test results (adapted from Harel and Zhou[23])*

	Liver pathology positive	Liver pathology negative	Liver pathology not performed
Hepatic scintigraphy positive	231	32	166
Hepatic scintigraphy negative	27	54	140
Total	258	86	306

TABLE 3 *Estimates with 95% CIs of sensitivity and specificity for different methods to adjust or impute missing data on reference standard outcome (based on data from* Table 2*)*

Procedure	Sensitivity		Specificity	
	Estimate	95% CI	Estimate	95% CI
Complete cases	0.895	0.858 to 0.932	0.628	0.526 to 0.730
Begg and Greenes[a]	0.836	0.788 to 0.884	0.738	0.662 to 0.815
Logit version Begg and Greenes[a]	0.836	0.835 to 0.838	0.738	0.735 to 0.741
A&C[b]	0.869	0.820 to 0.918	0.672	0.571 to 0.772
Rubin (logit)[b]	0.872	0.817 to 0.912	0.675	0.567 to 0.797
Wilson[b]	0.869	0.837 to 0.901	0.672	0.610 to 0.733
Jeffrey[b]	0.872	0.838 to 0.901	0.675	0.611 to 0.734
Z&L[b]	0.872	0 to 1	0.675	0 to 1

A&C, Agresti-Coull; Z&L, Zhou and Li method.
[a] Correction method.
[b] Multiple imputation method.

TABLE 4 *Diaphanography data: outcome of reference standard in those verified and fraction of unverified test results (adapted from Harel and Zhou 2006[23])*

	Breast cancer positive	Breast cancer negative	No verification
Diaphanography positive	26	11	30
Diaphanography negative	7	44	782
Total	33	55	812

TABLE 5 *Estimates with 95% CIs of sensitivity and specificity for different methods to adjust or impute missing data on reference standard outcome (based on data from* Table 4*)*

Procedure	Sensitivity		Specificity	
	Estimate	95% CI	Estimate	95% CI
Complete cases	0.788	0.649 to 0.927	0.800	0.694 to 0.906
Begg and Greenes[a]	0.280	0.127 to 0.434	0.974	0.960 to 0.989
Logit version Begg and Greenes[a]	0.280	0.275 to 0.285	0.974	0.973 to 0.975
A&C[b]	0.706	0.560 to 0.852	0.861	0.753 to 0.970
Rubin (logit)[b]	0.717	0.548 to 0.841	0.869	0.721 to 0.944
Wilson[b]	0.706	0.601 to 0.812	0.861	0.839 to 0.884
Jeffrey[b]	0.718	0.603 to 0.815	0.863	0.839 to 0.885
Z&L[b]	0.717	0 to 1	0.869	0 to 1

A&C, Agresti-Coull; Z&L, Zhou and Li method.
[a] Correction method.
[b] Multiple imputation method.

Complete case analysis seems to overestimate sensitivity and underestimate specificity, whereas Begg and Greenes's correction methods seem dramatically to underestimate sensitivity and overestimate specificity.

Harel and Zhou used the MAR assumption where missingness is related to index test results only. In practice, factors other than index test results, such as symptoms, signs, co-morbidity or other test results, may have driven the decision to apply or to withhold the reference standard. Imputation models can be extended to incorporate these additional sources of information, if known and collected, in an effort to improve the prediction of the reference standard outcome in unverified patients.

Correct imperfect reference standard

Description of the method

If a true gold standard does not exist, a logical next step is to search for the best available procedure to verify index test results. This is the rationale behind the concept of reference standard: the best available method to determine the presence or absence of the target condition, not necessarily without error. However, imperfection of the reference standard leads to bias known as reference standard bias.[11,19,53] If knowledge about the amount and type of errors of the imperfect reference standard is available, then this information can be used to correct estimates of diagnostic accuracy. We assume that all patients have received the imperfect reference standard, so there are no missing data in contrast to the correction methods described in the previous section.

Basic model

The basic model assumes that the error rates of the reference standard are known and that algebraic functions can be used to recalculate estimates of accuracy. Key publications by Hadgu and colleagues[54] and Staquet and colleagues[55] provide several algebraic functions, all based on the assumption of conditional independence between the index test and reference standard results. The assumption implies that the errors of both tests are independent of the underlying true status of the target condition, hence the index test and reference standard do not tend to err in the same patients.

Sometimes, the exact sensitivity and specificity of the imperfect reference standard are not known, but a range of plausible values are available. This range of values can then be applied in a sensitivity analyses that will produce a range of estimates of accuracy for the index test.

Extensions of the basic model

In many clinical situations, the assumption of conditional independence is unlikely to be true. Often the index test and reference standard have a tendency to make errors in the same (difficult) patients, particularly when index test and reference standard are methodologically related or measure the same physiologic alteration.[56] In the detection of cancer, for example, both the index test and reference standard may have difficulties in detecting early stages of cancer, and in clinical chemistry, contamination of body fluid samples may affect both the index test and reference standard. Algebraic functions have therefore been developed that build in the conditional dependence, such as the correlation in errors.[57] Unfortunately, the amount of correlation between errors is rarely known, hence this parameter is often varied over a wide range of plausible values.

Strengths and weaknesses

These correction methods can easily be applied using simple algebraic functions. The main limitation is that in most situations the true sensitivity and specificity of the imperfect reference standard are not known, nor does one know the amount of correlation among errors of the index test and the reference standard. If the chosen values do not match the true values, the resulting estimates of diagnostic test accuracy would still be biased or could become even more biased than the unadjusted ones.[58] For these reasons, researchers have tried to generalise the algebraic correction functions to avoid making untenable assumptions.[54] Latent class analysis subsequently evolved, which can be viewed as an extension of this method for situations where there is no information available on either the error rates of the reference standard or the true prevalence (see the section 'Latent class analysis', p. 26). In latent class analysis, the result of the (imperfect) reference standard is incorporated as just one source of information about the true disease status.

Applicability

The basic correction method can be applied in any situation if the following conditions are met: reliable information on the magnitude of the error rates of the reference standard is available and the conditional independence assumption is likely to be true, or there is reliable information about the

correlation between errors in the index test and the reference standard. In all other situations, the method is likely to render accuracy estimates that are biased, despite the use of a correction method.

Software

The algebraic functions can be used in widely available software programs such as Excel or any other statistical program.

Clinical example

The development of sensitive tests for bladder cancer is critical for its early detection. A gold standard does not exist for the evaluation of endoscopic tests as ideally this would necessitate removing the bladder for detailed pathological examination, a procedure that would not be ethically justifiable. Therefore, pathological evaluation is done only on biopsy or surgically removed tissue. In other words, only positive findings during endoscopy (index test) are verified. Histological examination alone is therefore not a reference standard providing perfect classification because of the unknown likelihood of missing cancerous lesions because they were not biopsied. This is the downside of not being able to verify negative findings of the index test.

Assumptions are necessary to correct for biases inherent in this approach. Schneeweiss and colleagues[59] provide an example of derivation of an interval of sensitivity estimates that includes the true value. They compared the ability of 5-aminolevulinic acid-induced fluorescence and white light endoscopy to detect bladder cancer. This is an example of a comparative diagnostic accuracy study with two index tests. They hypothesised that 5-aminolevulinic acid-induced fluorescence endoscopy is of superior diagnostic value. A total of 208 patients under surveillance after superficial bladder cancer were included. Multiple evaluations over time within the same patient were included, leading to a total of 328 endoscopic evaluations in which both procedures (index tests) were performed. They used sensitivity as the main accuracy parameter as the consequences of false negative cases would be worse and these results should be minimised by a good test.

The effect of having no gold standard is that there can be misclassification among the four cells in the 2-by-2 table. The observed true positive cell consists of lesions that are positive with the index test and confirmed on biopsy. It is unlikely that any observation in this cell should belong in another cell because we assume that cancer cells identified by pathological evaluation always indicate to cancer. However, in reality, there could be more observations in the true positive cell when some observations were wrongly classified as false positives, which can happen when not enough or not the correct biopsies were taken from a lesion. These observations would be classified as false-positives when in reality they belong in the true positive cell. The magnitude of this type of misclassification is assumed to be small but unknown.

For the false negative and true negative cells, the mechanism is analogous. If cancerous lesions are missed by both index tests, they are classified as true negatives, but in reality they are false negatives. False negatives can be observed in this study because of the paired design of this study; for example, patients underwent both index tests so that a true positive finding with one technique can be considered a false negative finding for the other technique if that technique missed that lesion. The magnitude of this misclassification of negative index test results is also assumed to be small but larger than the misclassification in the positive tests.

The following misclassification model was used to correct the counts in the observed 2-by-2 table and recalculate sensitivity based on these corrected counts. Let $p(\mathrm{A})$ be the probability of misclassifying true positives as false positive results and $p(\mathrm{B})$ be the probability of misclassifying false negative as true negative results. A more general equation can now be derived for estimating test sensitivity:

$$\text{Sensitivity} = \frac{\text{'true positive'} + p(\mathrm{A}) \times \text{'false positive'}}{\text{'true positive'} + p(\mathrm{A}) \times \text{'false positive'} + \text{'false negative'} + p(\mathrm{B}) \times \text{'true negative'}}$$

If it were assumed that there would be no misclassification, sensitivity could be easily calculated from the observed numbers in the 2-by-2 table, the so-called naive estimate. The analysis that Schneeweiss and colleagues propose comprises most optimistic, most pessimistic and some realistic assumptions to generate an interval of sensitivity.[59,60] The maximum interval of observable sensitivity for 5-aminolevulinic acid-induced fluorescence endoscopy ranged between 78 and 97.5%, and the best estimate for sensitivity based on realistic assumptions was 93.4% (95% CI 90 to 97.3). The best sensitivity estimate for white

light endoscopy was 46.7% (95% CI 39.4 to 54.3, maximum range 47.2–53%).

This method to determine the maximum possible range of sensitivity estimates in studies in which negative findings cannot all be verified is easily applied. Depending on the assumptions, a range of reasonable scenarios can be constructed and the corresponding sensitivities can be reported.

Construct reference standard

Differential verification

Description of the method

Instead of ignoring, imputing or correcting the non-verified results in the analysis, a second reference standard can be used to achieve complete verification. Use of different reference standards in this way is also known as differential verification (see also *Figure 2*, p. 4).[19] The second reference standard is usually a less invasive test and is less costly or less burdensome to patients. In the detection of pulmonary embolism (PE), for example, it is nowadays considered unethical to perform pulmonary angiography in patients with a low suspicion of PE and a negative D-dimer test result.[61] In many of these studies, only patients at high-risk for PE receive the best available reference standard (angiography), whereas 'low-risk' patients are likely to receive a different reference standard, such as clinical follow-up.

The use of different reference standards between patient groups is frequently encountered in the literature. In a survey of 31 diagnostic reviews, differential verification was present in 99 out of 487 (20%) primary diagnostic accuracy studies.[3] In most cases, incomplete verification is neither specified in the design nor completely at random. Triggers to perform the preferred reference standard in some patients and not in others include a positive result on the index test, positive results from other tests or the presence of risk factors for the condition of interest. This means that differential verification is selective and shows a non-random pattern, being based on decisions by the practitioner or patient.

Differential verification has been shown to lead to higher estimates of accuracy than studies using a single reference standard in all patients.[1,3] This type of bias is known as differential verification bias, or different reference standard bias, work-up bias or selection bias.[19,62] The effect of differential verification on estimates of accuracy is difficult to predict, as it depends on the proportion of patients verified differently, the selection process behind patients verified differently, the properties of the reference standards involved and their relation with the index test.[20]

Strengths and weaknesses

Differential verification appears to escape the bias of incomplete (partial) verification, but can lead to estimates of diagnostic accuracy that differ from those obtained with full verification by the preferred reference standard.[1,3] Moreover, estimates of sensitivity and specificity are more difficult to interpret, as they are based on multiple index test–reference standard combinations.[20] If complete verification by the preferred reference is not possible and different reference standards have to be used, the best approach is to incorporate differential verification in the design. This means prespecifying the group of patients that will receive the first reference standard and the group that will receive the second. An example could be that all patients with a positive index test are verified by one reference standard and all negative patients are verified by the second reference standard. In that case, the appropriate measures of accuracy are the positive and negative predictive values, calculated with corresponding reference standards. Because these measures are directly influenced by changes in prevalence the right design has to be chosen, such as cohort rather than case–control.

Field of application

Differential verification can be a reasonable option if several acceptable diagnostic tests are available that can serve as the reference standards. A prerequisite, however, is that the verification scheme is preplanned (design-based). As in any design, authors should report the rationale for each reference standard. Moreover, results should be reported separately for each index test–reference standard combination.

Software

Studies with differential verification produce standard accuracy results. No additional programming is necessary.

Clinical example

An illustrative example of differential verification that was (partly) design based can be found in the publication of Kline and colleagues.[63] The objective of the study was to evaluate the diagnostic accuracy of the combination of D-dimer assay with an alveolar dead-space measurement for rapid exclusion of pulmonary embolism.

In this multicentre study, V/Q scans, helical computed tomography, ultrasonography, clinical follow-up, angiography, death, events of deep venous thrombosis or new events of pulmonary embolism and combinations of these tests were used as diagnostic criteria to determine the presence or absence of pulmonary embolism. The authors describe a verification scheme to establish the final diagnosis that at first seems to be design based, but later on they state that the decision to order further imaging in patients with non-diagnostic V/Q scans was at the discretion of the attending physician.

The authors provide a table describing the different criteria for the presence of PE and the matching number of patients. Unfortunately, the corresponding results of the combined D-dimer–alveolar dead-space measurement were not stated in this table. In the results section, sensitivity, specificity, negative and positive likelihood ratios and the corresponding CIs were calculated in the usual way.

In the discussion, the authors do address the problem of differential verification. They state that computed tomography, for example, performs differently from angiography, and may have wrongly diagnosed some patients with PE who may have been free from PE.

The results of this study may not be very helpful to clinicians in practice, as it is unclear to which patients the estimates of diagnostic accuracy apply.

Discrepant analysis

Description of the method

Discrepant analysis, also referred to as discrepant resolution or discordant analysis, uses a combination of reference standards in a sequential manner to classify patients as having the target condition or not.[22,64,65] Discrepant analysis can provide estimates of prevalence and estimates of sensitivity and specificity of the index test without statistical modelling.

Initially, all patients are tested with the index test and one imperfect reference standard (*Figure 4*). Since the imperfect reference standard is known to be imperfect, the discordant or discrepant results (cases where index and first reference standard disagree) are retested (resolved) with an additional reference standard, frequently called the resolver test. The resolver test is usually a more invasive, more costly or otherwise burdening test, with better discriminatory properties than the first reference standard. The results of the resolver test are then used to update the final 2-by-2 table. Based on this final 2-by-2 table, estimates of accuracy for the index test are calculated.

Strengths and weaknesses

The method is straightforward and easy to carry out without statistical expertise. At first sight, this design seems to provide an efficient alternative to reduce the number of patients who have to be tested with the best available reference standard when this standard is either invasive or costly to apply. Yet a fundamental problem of discrepant analysis is that the verification pattern is dependent on the index test results.[64,66] Although discrepant analysis provides the status of the target condition for those who are retested, it does not provide that information for those not retested, which is usually the majority. Discrepant analysis therefore has the potential to lead to serious bias.[64,65,67–71]

The potential for bias has been demonstrated algebraically and numerically in the estimation of sensitivity and specificity in a situation where the resolver test is a perfect gold standard.[66,68,69] In this situation, discrepant analysis will lead to estimates of sensitivity and specificity for the index test that are biased upwards, so that the index test appears to be more accurate than it really is. The magnitude of this bias depends on the absolute number of errors of the index tests (false positive and false negative index test results) and the amount of correlation between the errors of the index test with the results of the first reference standard. These are correlated errors that will erroneously appear as either true positives or true negatives in the final 2-by-2 table as these patients will not enter the second stage and therefore will not be corrected by the perfect resolver test.

Green and colleagues showed that when the resolver test is not a perfect gold standard, even larger biases are possible, as the second imperfect reference standard can lead to further misclassification of the discordant results.[72] If errors of the stage 2 resolver test are correlated with errors of either the index test or the stage 1 imperfect reference standard, the diagnostic accuracy of the stage 2 reference standard will be biased in the assessment of discordant results. In other circumstances, the measured values may actually be closer to the true values.[66] In general, situations where the errors of the index test and the resolver test are related are of particular concern.

A method to correct for the potential bias in the discrepant analysis has been proposed,[66,73] but this

method is considered to be of limited applicability. The method requires that a random sample of patients with concordant results of the index test and the initial reference standard is tested by the resolver test. From this sample, the concordance rate of false results is estimated and the observed sensitivity and specificity are corrected accordingly. This procedure is adequate only if the resolver test is (nearly) perfect, a situation which occurs infrequently.

Field of application

Although discrepant analysis has been most frequently applied in the field of microbiology (detection of infections),[64,66] it can, in theory, be applied in all other medical fields. The general notice is to refrain from discrepant analysis because of the fundamental problem of the implicit incorporation of index test results in the definition of the true disease status leading to potential bias.[64,65] (http://www.fda.gov/cdrh/osb/guidance/1428.pdf). Only in special situations can the choice for discrepant analysis be defended if there is a perfect resolver test and a random sample of concordant results is also verified by the resolver test using the correction method proposed by Begg and Greenes.[22,48]

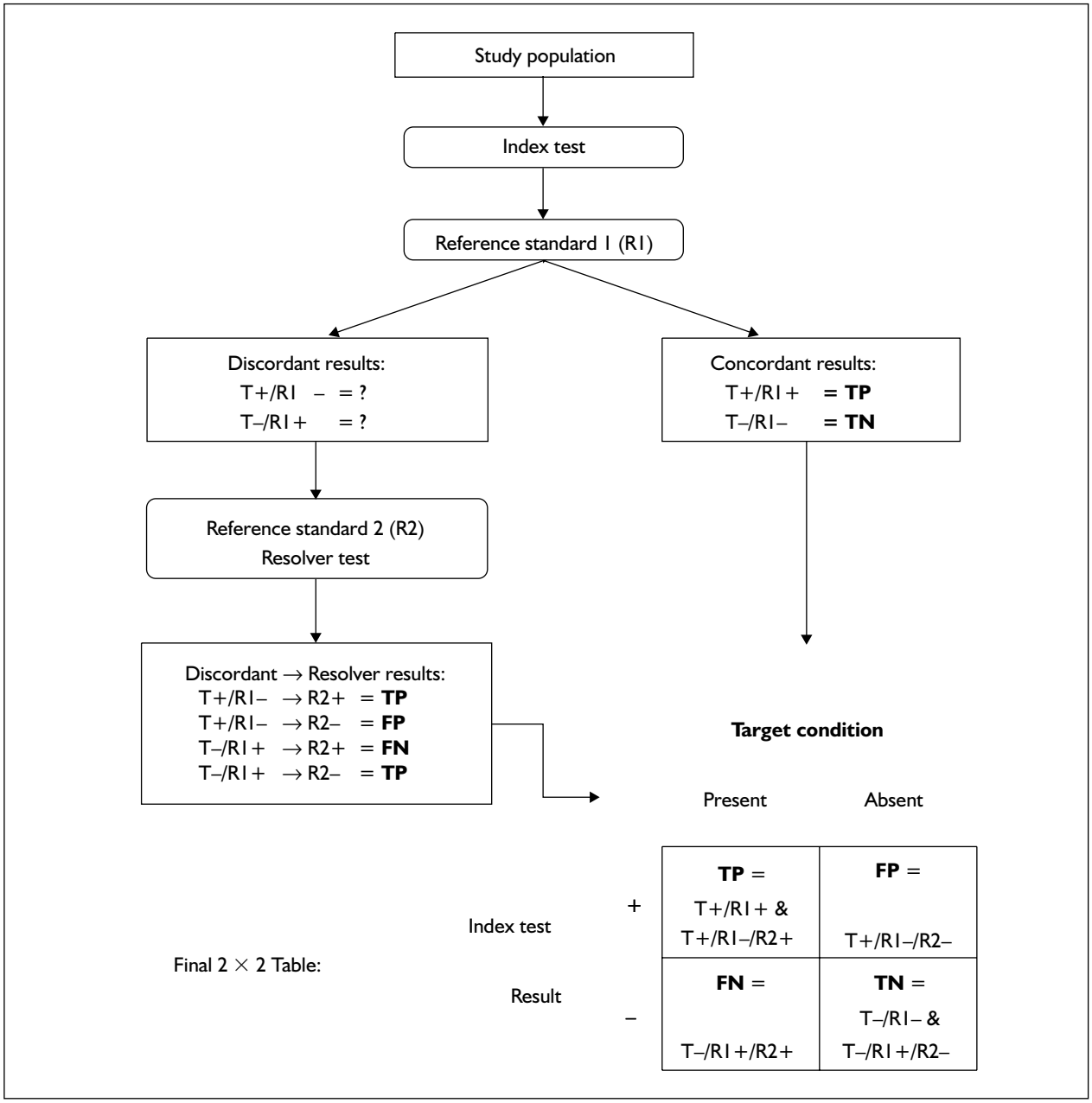


FIGURE 4 *Flow of patients and final 2-by-2 table in a study using discrepant analysis*

Software

Any statistical package can be used. The calculations can also be done with a pocket calculator.

Clinical example

To diagnose *Chlamydia trachomatis* infection, culture, DNA amplification methods such as polymerase chain reaction (PCR) and antigen detection methods such as enzyme immunoassay (EIA) are used (the same example is also used in composite reference standard and latent class analysis). Alonzo and Pepe[25] discuss the problems of assessing the accuracy of EIA (index test) as neither of the two other methods can be considered 'gold standards'. Culture is believed to have nearly perfect specificity, but also misses patients with *Chlamydia* infection (false negatives). PCR is believed to be more sensitive than culture to detect *Chlamydia trachomatis*. Alonzo and Pepe use the method of discrepant analysis to assess the accuracy of EIA by using culture as the initial reference test and PCR to resolve the discrepant results in the second stage (resolver test).[25]

The data are derived from the study of Wu and colleagues.[74] In the first stage, all 324 specimens are tested by EIA (index test) and culture (*Table 6*). In the second stage, only the discrepant results, for example, specimens which are positive by EIA and negative by culture ($n = 7$) or vice versa ($n = 3$), are retested with the resolver, PCR. The concordant results directly enter the final 2-by-2 table as true positives ($n = 20$) and true negatives ($n = 294$).

In the second stage, the 10 discordant results are tested with PCR and the outcome of the PCR test determines the classification in the final 2-by-2 table (*Table 7*). The EIA positive and culture negative results become either true positives if the PCR is positive ($n = 4$) or false positives if the PCR outcome is negative ($n = 3$). The same rule applies for the EIA negative and culture positive results ($n = 3$): the outcome of the PCR determines the final classification as either true negative if PCR is negative ($n = 1$) or false negative if PCR is positive ($n = 2$).

TABLE 6 *Initial classification using culture as the first (imperfect) reference standard*

	Culture	
EIA	+	–
EIA+	20	7
EIA–	3	294

Combining the results from the initial stage (*Table 6*) and the second resolver stage (*Table 7*) leads to the final 2-by-2 table (*Table 8*) on which the accuracy can be calculated.

The estimates of accuracy after retesting the 10 discordant results with PCR are as follows:

prevalence is $26/324 = 0.080$
sensitivity is $(20 + 4)/(20 + 4 + 3 - 1) = 24/26 = 0.923$
specificity is $(294 + 1)/(294 + 1 + 7 - 4) = 295/298 = 0.990$.

If we would have calculated the estimates of accuracy directly after using culture as a single reference standard, they would have been as follows:

prevalence is $23/324 = 0.071$
sensitivity is $20/23 = 0.870$
specificity is $294/301 = 0.977$.

Changes in this case are small because the frequency of discordant results was relatively low (3.1%). The validity of the approach depends on the error rate of the resolver test and the frequency of errors in the concordant results in the initial stage.

Composite reference standard

Description of the method

In the absence of a single gold standard, the results of several imperfect tests can be combined to create a composite reference standard. The results of the component tests of the composite

TABLE 7 *Results of the second stage: selective testing of discrepant results with PCR, the resolver test*

Discrepant results	Resolver test PCR	
	+	–
EIA+ and culture – ($n = 7$)	4	3
EIA– and culture + ($n = 3$)	2	1

TABLE 8 *Final classification based on discrepant analysis method*

EIA	Final classification	
	+	–
EIA+	20 + 4	3
EIA–	2	294 + 1

reference standard determine the presence of the target condition based on a prespecified rule. Such a composite reference standard is believed to have better discriminatory properties than each of the reference standards components in isolation.[24,25,64,66,75]

Basic model
A basic model is where two tests are applied in all patients and a prespecified (deterministic) rule is used to classify patients as having the target condition. Several definitions for the presence of the target condition can be used, depending on whether emphasis should be given to detecting or excluding the target condition and on the characteristics of the available reference tests. Most frequently, researchers define the target condition to be present if either one of the reference tests is positive. In the evaluation of the diagnostic accuracy of EIA in the detection of *Chlamydia trachomatis*, for example, a composite standard of two reference tests has been used: culture and PCR.[76] EIA and the two reference tests were applied in all patients and infection with *Chlamydia trachomatis* was diagnosed if either culture or PCR was positive. If both reference tests were negative, the person was labelled free of *Chlamydia trachomatis*.

It is important to note that the composite reference standard approach differs from discrepant analysis (see the section 'Discrepant analysis', p. 19), as the results of the index test themselves play no role in the verification procedure of the composite method. The difference between using a composite reference standard and differential verification is that in the composite method each patient receives all (necessary) components of the composite reference standards whereas in differential verification subgroups of patients are verified by one reference standard and other subgroups by a **different** reference standard (see also the section 'Differential verification', p. 18).

Extensions of the basic model
The composite reference standard approach can be extended to include more than two reference tests, as for instance in the detection of myocardial infarction, where chest pain, serological markers and electrocardiograms are used to define the presence or absence of the disease.[77] Increasing the number of reference tests amplifies the number of ways to define the presence of the target condition. If six reference tests are combined, for example, one possibility is the 'any test positive' definition of disease. Another definition was used in a study evaluating *Helicobacter pylori*, where a patient was regarded infected whenever two or more of the six available tests were positive.[78]

The efficiency of the composite reference standard can often be improved by avoiding redundant testing. If the classification rule is based on one of the two reference tests being positive, there is no need to perform a second reference test once the outcome of the first reference test is positive. Hence retesting is only necessary in patients with a negative result on the first reference test. Performing the more accurate reference test first lowers the number of patients needing additional testing.

Another extension is the use of statistical models to combine several reference test results to define disease status as in latent class analysis. This approach is discussed separately, as it does not use a prespecified deterministic rule for classifying patients having the target condition or not (see the section 'Latent class analysis', p. 26, for more details).

Strengths and weaknesses
The method is straightforward and easy to understand. It allows one to combine several sources of information in order to assess whether the target condition is present. The prespecified rule to define when the target condition is present increases transparency and avoids problems of incorporation and work-up bias. Unlike discrepant analysis, the application of the second reference standard is independent of the index test result.

A major issue is whether the combination of reference test results is a meaningful way of defining the target condition and whether residual misclassification is likely to be present. Do different reference standard tests look at the same target condition, but with different error rates, or do they define the target condition differently? The latter is thought to reduce the clinical usefulness.[24] The inclusion of more than two reference tests in the composite reference standard may then obscure the final definition of the disease.

A general problem is the determination of the 'cut-off value' to classify patients as having the target condition when several imperfect reference tests are available. Several suggestions have been made in literature. Alonzo and Pepe referred to latent class analysis as a tool for finding an optimal cut-off to define the target condition, but

TABLE 9 *Contingency table summarising composite reference for the* Chlamydia *data: EIA is the index test and the elements of the composite reference standard are culture and PCR.*

EIA	Culture		Composite reference standard: culture + PCR	
	+	–	+	–
EIA+	20	7	20 + 4	7 – 4
EIA–	3	294	3 + 2	294 – 2

they concluded against its use, as they see the model as being prone to bias due to its assumptions[64] (see also the 'Latent class analysis', p. 26). Some authors have suggested to use all possible 'cut-off values' to define the target condition rather than choosing a single one. They evaluate the index test against all these definitions and show how the accuracy of the index test varies across these definitions.[75,79]

Field of application

The composite reference standard method can be considered in all situations where a single gold standard does not exist, but several imperfect reference tests are available. One potential benefit is that it allows the researcher to 'adjust' the threshold for disease depending on the clinical problem at hand. In situations where missing the target condition outweighs the risk of wrongfully labelling persons as diseased, using a cut-off of 'any result positive indicates disease' can be advantageous. The composite reference standard method has been used in many areas, including infectious diseases[64,75,76] and gastroenterology.[79]

Software

The method is based on simple predefined classification rules, and requires no additional software. Measures of accuracy including CIs are calculated in their traditional way.

Clinical example

To diagnose *Chlamydia trachomatis* infection, culture, DNA amplification methods such as PCR and antigen detection methods such as EIA are used. (The same example is used in discrepant analysis and latent class analysis.) Alonzo and Pepe discuss the problems of assessing the accuracy of EIA (index test) as neither of the two other methods can be considered 'gold standards'.[25] Culture is believed to have nearly perfect specificity. PCR is believed to be more sensitive than culture to detect *Chlamydia trachotomatis*.

Alonzo and Pepe provide an example of assessment of EIA's accuracy using a composite reference standard based on culture and PCR.[25] They use the either positive rule to classify patients as having *Chlamydia* infection meaning that any specimen that is culture positive **or** PCR positive is composite reference standard positive and any specimen that is culture negative **and** PCR negative is composite reference standard negative. This approach allows one to use several sources of information in order to assess if an infection is present based on an *a priori* rule.

Because they use the either positive rule, there is no need to test all specimens with both reference tests of the composite reference standard: if a specimen is positive on the first reference test it will be positive on the composite reference standard and no further testing is therefore required. In the *Chlamydia* setting, in the first stage all specimens are tested by EIA (index test) and culture. In the second stage, only those specimens which are culture negative at the first stage are tested with PCR.

Considering the data from the study by Wu and colleagues[74] a contingency table summarising the two stages of composite reference standard is given in *Table 9*. After the first stage in assessing the performance of EIA using the composite reference standard, the 301 culture negative specimens (7 + 294) were tested with PCR. Six of these specimens were PCR positive and therefore considered to be infected based on the composite reference standard. The final estimates of accuracy are as follows:

prevalence is 29/324 = 0.090
sensitivity is (20 + 4)/(20 + 4 + 3 + 2) = 24/29 = 0.828
specificity is (294 – 2)/(294 – 2 + 7 – 4) = 292/295 = 0.990.

Because there is an *a priori* rule involving only the two reference test, the results of the index test (EIA) play no role in the classification of patients. This is different from the discrepant method because there the index test does play a role as only discrepant results move on to the second stage. To put it differently, in the composite

example the classification would not have changed if all specimens had been tested by both reference tests (culture and PCR). Because of the nature of the *a priori* rule (either positive), redundant testing can be avoided to gain efficiency.

One drawback of the composite reference standard is that it requires testing the typically large number of specimens negative by culture.

Panel or consensus diagnosis

Description of the method

In a panel or consensus diagnosis, a group of experts determines the presence or absence of the target condition in each patient based on multiple sources of information. These sources can include general patient characteristics, signs and symptoms from history and physical examination, and other test results. Information from clinical follow-up may also be included as an additional source of information to improve the classification of patients by the experts.

Variation exists in how to synthesise the input from different experts. Experts can discuss the information on each patient directly in a meeting and produce a consensus diagnosis using majority voting in case of disagreement. In other studies, experts determined the final diagnosis independently from each other and patients were only discussed in a consensus meeting in case of disagreement among experts. The final diagnoses from individual experts can also be combined using a statistical model.

Because a final diagnosis is determined in all patients, researchers can calculate estimates of accuracy in the traditional way without the use of specialised software.

Several practical decisions have to be made in setting-up a panel method. These include:

- the choice of experts: number and background
- the number of items to be considered in reaching a diagnosis and the way to present them
- whether or not to include the results of the index test
- combining the input from experts
- training and piloting.

Surprisingly little has been written in the medical literature about these practical decisions and how they can affect the final outcome. The discussion that follows is therefore largely based on theoretical considerations.

In the selection of experts, the qualities of the experts are expected to be of greater importance than the number of experts in a panel. Judgements by specialists may come closer to the true status of patients than a random or convenience sample of physicians. An uneven number of experts is sometimes recommended as it facilitates decision-making in case of majority voting.[80]

In general, all information relevant for the classification of patients is presented to the experts in a standardised way. Especially, subtle information from history taking might be difficult to get across on paper. Video footage of the original history taking or video films of dynamic radiological examinations might be an option.

Special care should be given to the role of the index test result in a panel diagnosis. If the index test result is given to the experts, its importance may be overestimated, leading to inflated measures of accuracy. This is known as incorporation bias.[2] On the other hand, the final classification in patients with and without the target condition may become better if the results of the index tests are disclosed to the experts. This would then reduce the problem of misclassification of the disease status. Most authors advise withholding the index test result from the experts. An attractive alternative is to use a staged approach where experts make a final diagnosis without the index test results first and then reveal the index test result and ask whether experts would like to change their classification.

As stated before, several methods of synthesising the input from different experts have been applied. Standard consensus meetings have been questioned because of problems arising from powerful personalities and also peer or group pressures. The Delphi procedure is a formal technique to collect and synthesise expert opinions anonymously to overcome these types of problems.[81]

Offering training and piloting sessions to experts can be used to increase agreement among experts. Piloting of the disease classification process may unearth problems that can then be addressed prior to the actual consensus process. Whether or not fixed decision rules need to be established during this piloting process is a matter of debate.

Strengths and weaknesses

Consensus diagnosis is an attractive alternative if a generally accepted reference standard does not

exist and multiple sources of information have to be interpreted in a judicious way to reach a diagnosis. Examples include target conditions that can lead to a wide range of symptoms and that cannot be diagnosed histologically. In those cases, the target condition can be defined by the clinicians, who take into account items of history, clinical examination, other test results and clinical follow-up, in order to determine clinical management. In such cases, the panel method closely reflects the clinically relevant concept of target condition. The accuracy statistics calculated in a study that relies on panel method to classify disease are likely to have high levels of generalisability to clinical practice.

Other advantages of this approach are the flexibility in determining conditions that have been poorly defined and the option of classifying conditions in a binary (target condition present or absent) and also in a multi-level way (as in severe, moderate, mild, no disease, indeterminate, etc.).

The downside of this method can be poor inter- and even intra-rater agreement, leading to discordance in disease classification between and within the experts. The levels of inter- and intra-rater agreements for the experts can be measured and should be reported in the study as part of validation of the consensus method.

Second, the subjectivity of the disease classification process may introduce bias. Incorporation bias or test review bias looms if the result of the index test is disclosed to the experts.

Third, the absence of a strict definition of disease in panel-based methods may be responsible for a divergence in the definition of 'disease' in the group of experts, who may have a different view of what constitutes the target condition of interest. Piloting and some form of standardisation can help in remedying this problem.

Fourth, the panel method can become a laborious enterprise if many items of information per patient need to be summarised and if many patients have to be discussed among several experts.

Finally, domination by assertive members of the panel can weaken any consensus procedure, although formal techniques such as the Delphi method are available to reduce this problem.

Application

The panel diagnosis method is well suited when there is no generally accepted reference standard procedure and multiple sources of information have to be interpreted in a judicious way to reach a diagnosis. In particular, the panel method is suited for target conditions that cannot be unequivocally defined. Examples in which panel diagnoses have been used include heart failure, the underlying causes in patients with syncope and underlying conditions in patients presenting with dyspnoea.

Software

No specialised software is needed, unless advanced procedures such as latent class analysis are used as a method of combining the results from various experts (for details, see the section 'Latent class analysis', p. 26).

Clinical example

Echocardiography is considered an imperfect reference standard test for heart failure. The European Society of Cardiology recommends that the diagnosis of heart failure be "based on the symptoms and clinical findings, supported by appropriate investigations such as electrocardiogram, chest X-ray, biomarkers and Doppler-echocardiography". Convening a consensus panel to establish the presence or absence of the target condition is then an appropriate approach to address the issue of imperfect reference standard in this study.

Rutten and colleagues evaluated which clinical variables provide diagnostic information in recognising heart failure in primary care patients with stable chronic obstructive pulmonary disease (COPD), and whether easily available tests provide added diagnostic information, in particular N-terminal pro-brain natriuretic peptide (NT-proBNP.[82] They studied patients over the age of 65 years with stable COPD diagnosed by a general practitioner without a previous diagnosis of heart failure made by a cardiologist. Clinical variables included history of ischaemic heart disease, cardiovascular medications, body mass index, displacement heart and heart rate. The tests included NT-proBNP, C-reactive protein, electrocardiography and chest radiography. An expert panel made up of two cardiologists, a pulmonologist and a GP determined the presence or absence of heart failure by consensus. The panel used all available information, including echocardiography, but did not use NT-proBNP in making the diagnosis. When there was no consensus, the majority decision was used to allocate the diagnostic category. Whenever a situation of evenly split votes arose, the majority decision amongst the two cardiologists and the GP

was used to reach a diagnosis, thus leaving out the vote of the pulmonologist. The situation of split vote only occurred in a minority of the cases (5/405, 1%). The authors re-presented a random sample of patients to the expert panel, blinded to the original decision, and found excellent level of reproducibility (Cohen's $k = 0.92$).

A limited number of items from history could help GPs identify those with heart failure. NT-proBNP and electrocardiography were the two tests that were found to be useful in improving the accuracy of diagnosis. As fully appreciated by the study authors, the study suffers from the possibility of incorporation bias, except in the case of the index test of NT-proBNP, as results of this test were not made available to the consensus panel. The authors postulated that the magnitude of the incorporation bias is likely to have been small as most diagnostic determinants were not crucial in the panel diagnosis process. This postulate is supported by the use of echocardiography, which is central in the diagnosis of heart failure, only as part of the reference standard, and not as one of the index tests.

No specific algorithms or guidelines were given in the article on how the panel weighted and combined the various items of history, examination and investigations findings, although references were made to previous studies which may have given some guidance on this. Standardisation of the disease categorisation process and threshold for disease positivity are recognised to be problems with heart failure. It is not clear if domination by a specific person or persons in the consensus panel was an issue, as this is certainly plausible when the panel has generalists and specialists with varying levels of expertise in cardiology. Despite these weaknesses, this study is an excellent example of the use of a consensus panel as a reference standard.

Latent class analysis

This group contains many variations, but all methods have in common the use of a statistical model to combine different pieces of information (test results) from each patient to construct a reference standard. These methods acknowledge that there is no gold standard and that the available tests are all related to the unknown true status: target condition present or absent.[64,83–89]

The problem that the outcome of interest cannot be measured directly occurs in many research situations. Examples include constructs such as intelligence, personality traits or, as in our case, the true diagnosis. These unobservable outcomes are named latent variables. These latent variables can only be measured indirectly by eliciting responses that are related to the construct of interest. These measurable responses are called indicators or manifest variables. Latent variable models are a group of methods that use the information from the manifest variables to identify subtypes of cases defined by the latent variable.

The problem of evaluating an index test in the absence of a gold standard can be viewed as a latent variable problem. In our case we have dichotomous latent variable, namely whether or not patients have the target condition. When the latent variable is categorical (dichotomous being a special case within this group), the latent models are referred to as latent class models, whereas when the latent variable is continuous they are named latent trait models. Results of the index test and other imperfect tests are then the observable (manifest) variables that can be used to estimate the parameters that are linked to true diseases status, like sensitivity, specificity and prevalence. Maximum likelihood methods can be used to estimate these parameters of the latent model.

The latent class approach has similarities with the panel or consensus diagnosis methods, which also uses multiple pieces of information to construct a reference standard. In the panel method, experts determine whether each patient has the target condition given a set of test results, whereas a formal statistical model is used to obtain the statistics of interest in the latent class approach.

Basic model

The statistical framework underlying latent class models can be illustrated using the following basic example. In this example we have three different tests being applied in all patients with each test producing a dichotomous test result (e.g the test is either positive or negative). All three tests relate to the same target condition, but none of them is error free. For a single test, the probability of obtaining a positive test result can be written as the sum of finding a positive test in a patient who has the target condition (true positive result) or a positive test result in a patient without the target condition (false positive result). These probabilities (Prob) can be written as a function of the following unknown measures: prevalence (prev), sensitivity of test 1 ($sens_1$) and specificity of test 1 ($spec_1$). The probability of finding a true positive result for test 1 (TP_1) can be written as

$$\text{Prob}(TP_1) = \text{prev} \times \text{sens}_1$$

and the probability for obtaining a false positive result (FP) as

$$\text{Prob}(FP_1) = (1 - \text{prev}) \times (1 - \text{spec}_1)$$

Therefore, the probability of a positive test results for test 1 is the sum of these two probabilities

$$\text{Prob}(+) = \text{Prob}(TP_1) + \text{Prob}(FP_1) = \text{prev} \times \text{sens}_1 + (1 - \text{prev}) \times (1 - \text{spec}_1) \quad (1)$$

In the same way, we can write the probability for obtaining a true negative (TN) result as

$$\text{Prob}(TN_1) = (1 - \text{prev}) \times \text{spec}_1$$

and for a false negative (FN) test result as

$$\text{Prob}(FN_1) = \text{prev} \times (1 - \text{sens}_1)$$

and the probability of a negative test result as

$$\text{Prob}(-) = \text{Prob}(TN_1) + \text{Prob}(FN_1) = (1 - \text{prev}) \times \text{spec}_1 + \text{prev} \times (1 - \text{sens}_1) \quad (2)$$

Of course, the probabilities of the other tests are defined in the same way, but then using the sensitivity and specificity of that specific test. This means that there are seven unknown parameters in this example: one prevalence parameter and the sensitivity and specificity for each of the three tests (1 + 6 = 7).

With three different dichotomous tests, there are eight possible combinations of test results: all tests being positive, three variations where two tests are positive and one negative, three situations where one test is positive and two negative, and the situation where all tests are negative. By using the probabilities for a positive [equation (1)] or negative test result [equation (2)] for each test, we can write down the likelihood of observing each pattern of test results. Under the assumption of statistical independence, the likelihood of observing a specific pattern can be written as the probability of observing that pattern in patients who have the target condition plus the probability of observing the same pattern in patients without the target condition.

For instance, the probability of observing the pattern ++– is

$$\text{Prob}(++-) = \text{sens}_1 \times \text{sens}_2 \times (1 - \text{sens}_3) \times \text{prev} + (1 - \text{spec}_1) \times (1 - \text{spec}_2) \times \text{spec}_3) \times (1 - \text{prev})$$

When we carry out such a study, we would observe the number of patients for each of the eight patterns of test results. The sum of these numbers has to be equal to the total number of patients in the study, which means that we have 8 – 1 = 7 degrees of freedom. If the degrees of freedom are equal to or greater than the number of parameters to be estimated, standard maximum likelihood methods can be used to obtain a (unique) solution. More mathematical details can be found elsewhere.[85,90]

Extensions of the basic latent class model

Several extensions to this basic model (dichotomous latent variable, three dichotomous tests assuming uncorrelated errors) have been formulated. Here we discuss the main extensions relevant for diagnostic research.

The number and type of tests

Latent class models are flexible and can incorporate dichotomous results, but also ordinal test results or continuous test results.[90] The model can easily be extended to incorporate the results of more than three tests. Including additional tests is beneficial from a modelling point of view as it increases the available degrees of freedom (more tests lead to more test results combinations). More degrees of freedom mean that more parameters can be estimated, for instance a correlation parameter to acknowledge that errors between tests might be correlated (see also the section below on conditional dependence). The extra degrees of freedom can also be used for additional checks of the fit of the model.

Much attention has been given to estimating the accuracy of a test when only one other additional test (e.g. imperfect reference standard) is available.[91,92] In this case, the number of parameters to estimate (1 prevalence + 2 sensitivities + 2 specificities = 5 unknown parameters) is larger than the available degrees of freedom (4 test results combinations: ++, +–, –+, –– minus 1 = 3 degrees of freedom). This means that no optimal maximum likelihood solution can be identified: different combinations of values of prevalence, sensitivities and specificities fit the data equally well. Only through restrictions can we estimate the parameters of the model, for instance assuming that the sensitivities and specificities of the two tests are equal. Another option is to incorporate prior information about the parameters into the model by using a Bayesian approach to estimate their values (see the section 'Bayesian framework', p. 28).

Conditional dependence

The basic model assumed that the results of the three available tests were independent conditional

on the true disease status, also known as local independence. Independence in this context means that the errors of the test are not correlated, e.g. if a diseased patient is misclassified by one test, it does not increase the likelihood that this patient will be misclassified by another test. In other words, there is no group of 'difficult' patients in whom several tests perform less than expected. This assumption might hold if tests measure different manifestations of the target condition and/or use different clinical methods. In the detection of *Chlamydia*, for example, when antigen detection with EIA, cell culture and DNA amplification with PCR is used, it is less likely that these tests make the same type of errors than if two of the three tests were DNA amplication tests. An example where the assumption of conditional independence is likely to be violated is in the detection of lumbar herniation if all three available tests are imaging tests, such as magnetic resonance imaging (MRI), CT and radiography, all focusing at the detection of visible abnormalities of the discus.

For many situations, the independence assumption is unlikely to be true. Ignoring the correlation of errors between tests can seriously affect the estimates of accuracy.[26,64] To solve this problem, we can incorporate the correlation of errors between tests into the model.[26,88] This requires the estimation of additional parameters and therefore more degrees of freedom to estimate them. Additional degrees of freedom can be obtained by including more tests, repeating the study in another population with a different prevalence of the condition (but unchanged accuracy estimates) or incorporating prior information using a Bayesian framework (see the next section).

Bayesian framework

The parameters of a latent class model can also be estimated through a Bayesian approach instead of the more traditional maximum likelihood estimation (frequentist approach). The Bayesian framework estimates the same latent model and parameters, but it explicitly incorporates prior information to the model.[54,91,93–95] In the Bayesian approach, the unknown parameters are all treated as random variables having a probability distribution. Information available on each parameter prior to collecting the data is summarised in the prior probability distribution, which is then combined with information from the observed data to obtain a posterior probability distribution for each parameter. The posterior distribution can be used to obtain point estimates for the mean and median sensitivity and specificity with credibility intervals, which can be loosely interpreted as Bayesian CIs.

Prior distributions typically are uninformative or determined from the published literature or in consultation with experts. The Bayesian approach is sensitive to the chosen prior distribution used: using different priors can lead to differences in estimates of diagnostic accuracy.

The Bayesian approach can be particular helpful in ill-defined situations, such as situations where the number of parameters to be estimated is large relative to the available degrees of freedom. The use of prior information in combination with simulations means that the Bayesian approach can obtain estimates where traditional methods fail.[96] Bayesian modelling is more complicated as it requires programming, simulations and validations of the results.

Strengths and weaknesses

The latent class analysis methods are well documented, statistically sound and have been applied in many areas of research.[90] The characteristics of this method have been extensively studied, including in simulations studies[64,85,88] and in studies that compare latent class estimates of accuracy with the estimates derived from a classic design where a gold standard was available.[97] The model provides estimates of sensitivity and specificity for all tests incorporated in the model, which are the indexes most commonly used in test evaluation research. Latent class analysis is a flexible approach that can incorporate different types of test results (dichotomous, ordinal and continuous).

The drawbacks of latent class analysis are well described in the literature.[64,85,89] The greatest concern is not related to statistical issues but to a more basic principle. In a latent class analysis, the target condition is not defined in a clinical sense. Because we make no clinical definition of disease in latent class analysis, clinicians can feel uncomfortable about what the results represent.[24,64,85]

Latent class analysis does not fully comply with a basic principle of test evaluation research. When evaluating the performance of an index test, it needs to be compared with a standard that is independent of the index test itself. In latent class analysis, however, this principle is partly violated because the index test results are often used to construct the reference standard. Considerations similar to those mentioned in the panel diagnosis

methods apply (see the section 'Panel or consensus diagnosis', p. 24). This problem lies more in the panel method because experts may overestimate the capabilities of the index test, whereas in the latent class setting a formal statistical model is applied, which is not sensitive like humans to the reputation of any test.

Like any statistical model, the validity of the results of a latent class model depends on whether the underlying assumptions are met. Whether or not a latent class model assumes conditional independence is a critical issue. Simulation studies have shown that violation of conditional independence biases the estimates of diagnostic accuracy.[26,64] If tests are positively related to diseased and/or not diseased patients, latent class analysis will yield accuracy estimates which are too high, especially for the more accurate test. A problem is that the conditional independence assumption cannot be fully tested in a model with three dichotomous tests. Additional or repeated testing to increase the degrees of freedom might not be feasible for ethical or economic reasons.

These observations have led to caution against the use of latent class analysis in practice when only three tests are available. Some even extrapolate this view to all latent class models, as even when it is possible to incorporate more information in the model, unverifiable assumptions about dependence are still required.[64,85]

Field of application

Latent class analysis can be applied in those situations where multiple pieces of information are available for each patient.

Software

Several more advanced statistical packages have implemented latent class analysis, such as Splus and R. In addition, there are programs specially developed for latent variable models such as Latent Gold and LEM. Free software to perform Bayesian latent class analysis is available, requiring only a user registration (WinBugs).

Clinical example

In the detection of *Chlamydia trachomatis*, culture, DNA amplification methods such as PCR and antigen detection methods such as EIA have been modelled with latent class analysis (the same example is also used in discrepant analysis and latent class analysis). Culture is believed to have nearly perfect specificity. PCR and EIA are believed to be more sensitive than culture to detect *Chlamydia trachomatis*. In *Table 10*, we present the observed test results of these tests in a group of 324 persons, as reported by Alonzo and Pepe.[64] The frequency of occurrence for each pattern of test results is given. Latent class analysis uses the information displayed in the table to determine associations between the diagnostic tests. More detail on the analytical expressions for estimates can be found in the paper by Pepe and Janes.[85] Referring to the detection of *Chlamydia trachomatis* with cell culture as the imperfect reference standard, specificity is assumed to be close to 100%, whereas a wide range of values have been reported for its sensitivity.[93,98,99] Different priors can be incorporated in the Bayesian approach, for instance an optimistic prior (using a uniform distribution in the range 80–90%) and a pessimistic prior (in the range 55–65%) for the sensitivity of culture. If nothing is known about the index test, a prior distribution for sensitivity and specificity allowing equal weighting in the 0–100% range can be used.

TABLE 10 *Pattern of tests results and their frequency of occurrence*

Pattern	PCR	EIA	Culture	Observed frequency
1	+	+	+	19
2	+	+	–	4
3	+	–	+	3
4	–	+	+	1
5	+	–	–	2
6	–	+	–	3
7	–	–	+	0
8	–	–	–	292

Parameter estimates derived from latent class analysis are as follows:

Prevalence:	0.081
Sensitivity and specificity of EIA:	0.909 and 0.990
Sensitivity and specificity of culture:	0.834 and 0.997
Sensitivity and specificity of PCR:	1.000 and 0.995

Validate index test results

In all previous methods, the focus was on correcting imperfect reference standards, or constructing a new or better reference standard, to calculate the classical indexes of diagnostic test accuracy such as sensitivity and specificity. In this section, we discuss approaches to evaluate index

test results that go beyond the diagnostic accuracy paradigm. Validation is an alternative process to evaluate a medical test in the absence of an unproblematic and equivocal reference standard. In this context, validity refers to how well the index test measures what it is supposed to be measuring.[100–102]

Many instruments in the social sciences and elsewhere in science rely on a validation process to determine whether or not the instrument can serve its purpose. In these cases, the diagnostic accuracy paradigm often cannot be used because the 'truth' cannot be observed. The latter applies to many hypothetical or conceptual constructs. We cannot observe a construct, or any latent variable (see the section 'Latent class analysis', p. 26). What we can do is to observe associated attributes, as, for instance, sweating, moaning and asking for pain medication in the evaluation of pain.[101,103] In a similar way, we can evaluate associations between the index test and these attributes.

Three different forms of validity have traditionally been distinguished: content validity, criterion-related validity and construct validity.[101] In this context the classical diagnostic accuracy design (*Figure 1*, p. 2) refers to criterion-related validity where the reference standard provides the criterion against which the index test is validated. Several other definitions have been provided, including intrinsic validity, logical and empirical validity, factorial validity and face validity. This proliferation is in part based on a misinterpretation of the classical paper by Cronbach and Meehl, who pointed to the quintessential nature of **construct validation** in all areas where criterion-oriented definitions cannot be used.[104]

Approaches towards the validation of medical tests explore meaningful relations between index test results and other test results or clinical characteristics, none of which can be uplifted to the status of a reference standard, either isolated or in combination. Relevant items can come from the patients' history, clinical examination, imaging, laboratory or function tests, severity scores and prognostic information.

Construct validity is not determined by a single statistic, but by a body of research that demonstrates the relationship between the test and the target condition that it is intended to identify or characterise. Validating a test can then be understood as a gradual process whereby we determine the degree of confidence we can place on inferences about the target condition in tested patients, based on their index test results. That degree of confidence is based on a network of associations between the test results and other pieces of information in tested patients.

One important way to validate an index test is to use dedicated follow-up to capture clinical events of interest in relation to index test results. If the index test will be used for predicting future events, such a prognostic study can be seen as an evaluation of the predictive validity of the test.[101] The question of whether it is safe to withhold further testing in patients with low probability of pulmonary embolism and a negative D-dimer result has been studied by looking at the 3-month incidence of venous thromboembolism, and reported as such.[61]

One step further is the evaluation of a test within randomised clinical trials of therapy. One aim of the test is to identify patients who are more likely to benefit from the new, active treatment.[105,106] A possible design corresponding to that study question is displayed in *Figure 5*.

Note that all patients receive the index test, but that none of them receives a reference standard. After testing, patients are randomly allocated to either treatment or clinical follow-up. At the end of follow-up, the data collected can be displayed in four 2-by-2 tables (*Figure 5*). Different study objectives can be answered by these tables. In Tables A and B in *Figure 5*, a conclusion can be drawn as to whether treatment is beneficial compared with clinical follow-up in index test positive patients and test negative patients, respectively. Odds ratios and CIs can be calculated to underpin the conclusions. In Tables C and D, conclusions can be drawn concerning the prognostic value of the test within the context of subsequent clinical decision-making. Table C refers to the ability of the index test to identify patients who are likely to benefit from therapy. The test properties can be either expressed as event rates (of, for example, poor events) or odds ratios. Table D refers to the ability to discriminate between different risk categories for a specific event. Several modifications of the basic randomised controlled trial (RCT) model to evaluate tests have been described by Lijmer and Bossuyt.[105]

Basic design

No single basic design exists for construct validation studies. There are many possible designs to evaluate the validity of an index, as there are many different predictions possible based on our theory or construct.

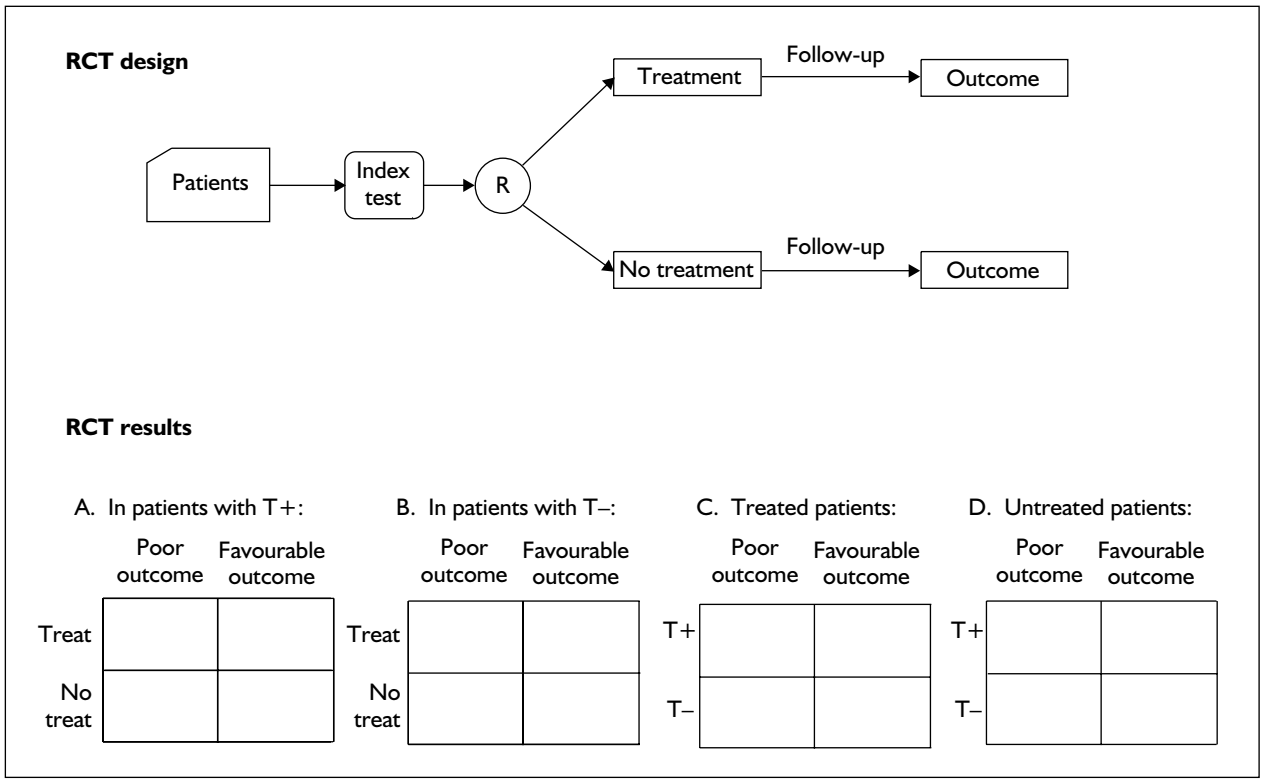


FIGURE 5 *Randomised controlled trial (RCT) using index test results as a prognostic marker (T = test)*

As an example, we look at the evaluation of a new questionnaire to evaluate psychological stress. Many hypotheses may be generated, including that level of stress increases heart frequency, and also blood values of cortisol. An association study could be conducted to evaluate the correlation between the score, heart frequency, blood levels of cortisol and past exposure to psychological stress, such as surviving a Tsunami. An additional approach would be to perform a factor analysis, to evaluate whether the elements of the questionnaire belong to one or more dimensions (factors), and to match if the identified factors correspond with the theory of psychological stress. Still another theory could point out that stress is artificially induced by an intervention, such as showing pictures of tortured persons. If so, participants should have higher scores on the questionnaire after such an intervention. Stated alternatively: if patients with high scores are given a tranquilliser, then their scores should drop. Here an intervention study would be conducted. If scores drop after tranquilliser intake in those with initially high scores, whereas those with low scores remain at approximately the same level, then the new questionnaire can be considered to have construct validity. Additional studies would have to be conducted. Gradually, our confidence in the construct validity and the inferences we can draw from measuring that construct will grow.

Several ways exist to express associations between the measured attribute and other attributes, events or subgroups, including correlation indexes, odds and risk ratios and absolute and relative differences.

Strengths and weaknesses

In situations where no reference standard is available, association studies may be the only type of study that can be performed. The validation of the index test will depend on the underlying theories about the target condition, as the latter provides us with hypotheses about the kind of associations between index test and attributes that have to be evaluated. If the theory is wrong, totally or in part, any quantitative expression of the strength of the association may be misleading.

Whenever the index test results fail to show the hypothesised network of associations between the target condition and other observations, more than one conclusion is possible: the index test has low validity, the theory about the target condition is not correct or both the index test and the theory are inadequate.

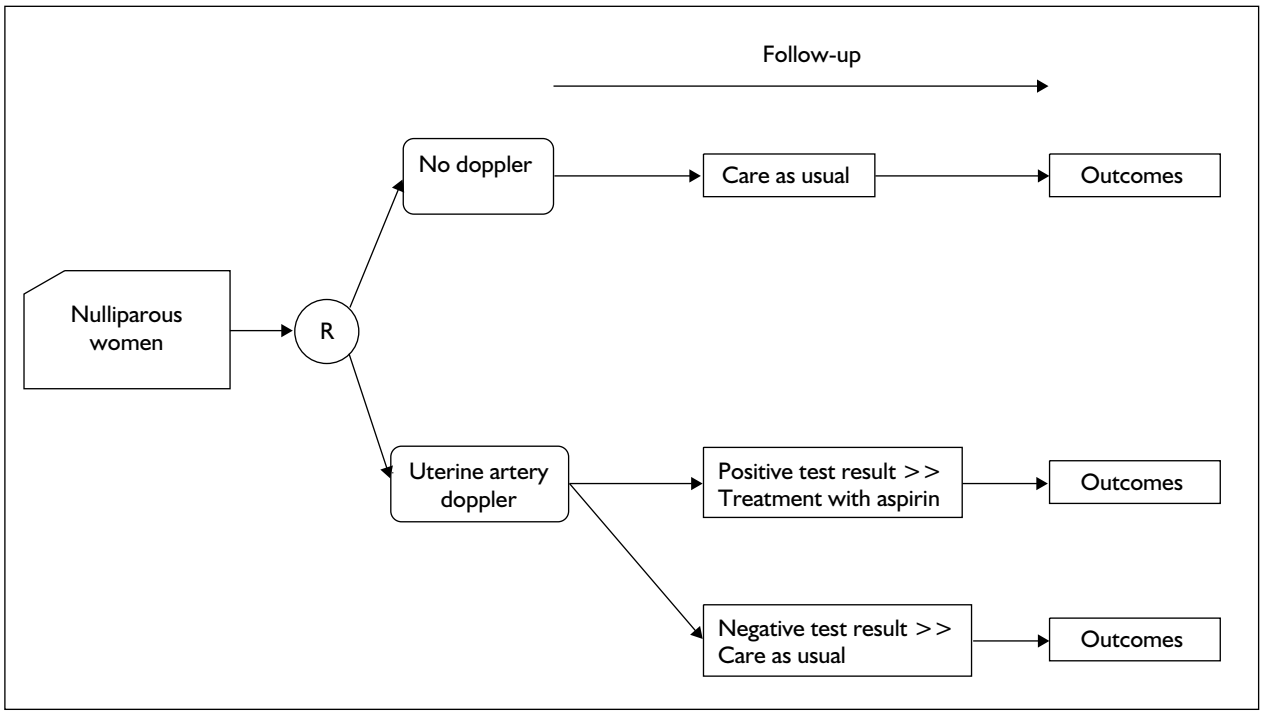


FIGURE 6 *Design of the uterine artery Doppler trial*

Specific problems related to clinical follow-up have been identified. The nature of the target condition may change during clinical follow-up. Even when present, the target condition is not guaranteed to produce detectable events or complaints. The length of follow-up has to be chosen judiciously for each specific target condition.

An additional problem arises when an intervention is used. If the intervention is chosen badly, and does not achieve what it was intended to achieve, the construct validity of the index test will be masked. If the outcome is favourable, any chance cannot be attributed exclusively to the discriminatory characteristics of the index test. Hence test evaluation and management consequences cannot be disentangled in this type of test evaluation.

A major drawback of construct validity is that there is no single type of study that can be prescribed, as in diagnostic accuracy studies. Several studies have to be performed, before enough confidence in the validity of the index test can be achieved.[107]

Field of application

We hypothesise that this type of test evaluation can be done in all areas of medicine where an accepted reference standard, or any other criterion to determine the diagnostic accuracy of a test, is absent. In the discussion (Chapter 5), we will argue that validation is a more universal way of appraising the value of medical tests.

Clinical example 1: randomised controlled trial

We will use a study by Subtil and colleagues titled 'Randomised comparison of uterine artery Doppler and aspirin (100 mg) with placebo in nulliparous women' to explain some of the design and efficiency issues that play a role when using an RCT to evaluate a test.[108]

The objective of this study was to assess the effectiveness of a strategy of pre-eclampsia prevention based on routine uterine artery Doppler examination during the second trimester of pregnancy, followed by a prescription of 100 mg of aspirin in those with abnormal Doppler, versus no Doppler strategy. The population were nulliparous women (no previous delivery before ⩾22 weeks) between 14 and 20 weeks' gestation, with no history of hypertension. Randomisation was used to determine whether women would receive uterine artery Doppler testing between 22 and 24 weeks or not. Women in the no testing arm of the trial would receive usual care, as would the women in the testing arm with a negative test result (for the design, see *Figure 6*). The outcomes of the trial were the development of pre-eclampsia, having a baby small for gestational age and perinatal deaths. The trial did not find a difference

in outcomes between the tested (Doppler group) and the non-tested group (no Doppler).

This design offers an assessment of the Doppler test and aspirin therapy combination, as women are randomised to Doppler or no Doppler groups, with treatment of the Doppler positive women in the Doppler arm. It addresses the question: does testing with Doppler (plus aspirin for test positive cases) prevent pre-eclampsia? In this design, if testing and treatment were shown not to be effective, it could be because the test is inaccurate or because aspirin is ineffective. This design, therefore, addresses two questions in one, but unfortunately, it is not often possible to segregate the accuracy of the test from the effectiveness of the treatment.

This design has various weaknesses. First, the design is not efficient and requires large samples to achieve satisfactory power. This is because the only women contributing to the expected difference in outcome in the two trial arms are those who belong to a small subgroup of those with abnormal Doppler result in the Doppler arm. In this study, 2491 women were randomised in a 2:1 ratio to Doppler and control arms, and of these 2491 women, an abnormal Doppler result was found in 239 (10%) women and these received aspirin. Even with a 2:1 ratio of randomisation to improve the numbers of women who had abnormal Doppler result (and, therefore, received the active treatment, aspirin), the total number of women who contributed to the difference between the Doppler and control arms was just 239, suggesting an under-powered study.

Even if this trial recruited many more thousands of women, and became adequately powered, and a benefit is subsequently shown for the Doppler arm, this may not be a vindication of benefit for the combination of Doppler test and aspirin therapy. This is because more women in the Doppler arm would have received aspirin compared with the no Doppler arm, and as aspirin has been shown to be generally effective in reducing pre-eclampsia,[109] then it is possible that the Doppler arm would have shown benefit regardless of whether the Doppler test was a good predictor of pre-eclampsia or not or, indeed, whether the Doppler test was done or not. This is because about 10% of women would have received aspirin in the Doppler group compared with none or few in the control group.

This design is, therefore, of limited use in assessing the role of uterine Doppler testing for aspirin therapy.

Clinical example 2 and 3: validation studies other than trials

An example of a validation approach can be found in a series of studies that evaluated the use of troponin to identify acute coronary syndrome in chest pain patients. Initial studies have evaluated cardiac troponin in an accuracy framework, using the original WHO definition of acute myocardial infarction. The latter invites the use of a composite reference standard, looking at characteristic ECG changes – either ST segment elevation or the development of new Q waves – confirmed by significant changes in serial cardiac enzymes.[110] Other studies have looked at 30-day outcomes in patients admitted without ST segment elevation on the initial ECG. They found that cardiac events increased significantly with increasing cardiac troponin values, suggesting that the degree of troponin elevation, and also clinical variables such as previous myocardial infarction and an ischaemic ECG, should be considered when deciding treatment.[111] These and other studies have led to the recommendation that cardiac troponins should be the preferred markers for the diagnosis of myocardial injury.[112] More recent investigations have indicated that increases in biomarkers upstream from markers of necrosis, such as inflammatory cytokines, cellular adhesion molecules, acute-phase reactants, plaque destabilisation and rupture biomarkers, biomarkers of ischaemia and biomarkers of myocardial stretch may provide an even earlier assessment of overall patient risk. The studies to support this claim are all closer to the validation framework than to the accuracy paradigm.[113]

Another example of association studies for test validation includes the evaluation of new tests for latent tuberculosis infection (LTBI), a condition for which no gold standard exists. The tuberculin skin test (TST) has been the standard test to detect LTBI, although it is known to have false results, particularly in people with previous Bacillus Calmette–Guérin (BCG) vaccination. Interferon-gamma assays (e.g. Quantiferon and Elispot) have recently been developed for use as new tests for LTBI. Tuberculosis infection evokes a strong T-helper 1 (Th1) type cell-mediated immune response with release of interferon-gamma. *In vitro* assessment of interferon-gamma production in response to mycobacterial antigens can be used to detect latent infection with *Mycobacterium tuberculosis*. Blood infected with *M. tuberculosis* contains a specific clone of T-lymphocytes stimulated by exposure to the ESAT-6 or CFP-10 antigen. The Elispot assay is based on the detection of interferon-gamma, which these

ESAT-6 and CFP-10 specific T lymphocytes secrete. As there is no gold standard or a suitable reference standard for LTBI against which the comparative accuracy of the TST and Elispot could be assessed, a validation study described below provides a good alternative to the classical diagnostic accuracy paradigm. The risk of infection is greatest among those contacts who share a room with the index case for the greatest length of time. This means that airborne transmission increases with length of exposure and proximity to an infectious case of tuberculosis. Hence results of tests for LTBI should correlate with level of exposure, and the test with the strongest association with exposure would be likely to be the most accurate. This issue can be evaluated in observational studies that ascertain exposure to tuberculosis in a relevant population and setting (e.g. outbreak investigation), perform various index tests of interest in all eligible subjects and compare test results with exposure status.[114] TST and Elispot response to ESAT-6 and CFP-10 were evaluated in this manner in 535 subjects during a school outbreak.[114] In this outbreak investigation, four exposure groups based on proximity and shared activities with an infected case were established following detailed interviews undertaken by a school nurse. The association of index test with exposure status was examined using the gradient of dose–response relating test results to degree of tuberculosis exposure. Comparison of these gradients showed that Elispot test was statistically significantly better than TST. Elispot correlated significantly more closely with *M. tuberculosis* exposure than did TST on the basis of measures of proximity ($p = 0.03$) and duration of exposure ($p = 0.007$) to the infected case. TST was significantly more likely to be positive in BCG-vaccinated than in non-vaccinated students ($p = 0.002$), whereas Elispot results were not associated with BCG vaccination ($p = 0.44$). The authors concluded that Elispot offers a more accurate approach than TST for identification of individuals who have LTBI. Further validation may come from observational studies evaluating whether new test results are predictive of development of active tuberculosis in the future.

Chapter 5

Guidance and discussion

In this report, we have summarised a number of methods that can be used to evaluate tests in the absence of a gold standard, an error-free method to establish with certainty the presence or absence of the target condition in tested patients. In this chapter we present our conclusions as guidance to researchers and emphasise the need to enrich or enlarge the accuracy paradigm in the evaluation of medical tests.

Guidance for researchers

The guidance we have distilled from the existing evidence and the recommendations from researchers can be summarised as follows (*Figure 7*). Readers should be aware that this flowchart can only provide general guidance because many factors have to be balanced when choosing one method over another.

The first question to be answered is whether the uncertainty about the test can be satisfactorily answered by knowing its accuracy. The diagnostic accuracy of a test is an expression of how well the test is able to identify tested persons with the target condition. The limits of this report do not allow us to discuss at length the various other ways in which a test can be evaluated, including randomised trials of test–treatment combinations that can document whether the use of a test will lead to improved patient outcomes.[106] We also omitted from consideration many of the lower-level evaluations that usually precede evaluations of diagnostic accuracy, including studies examining the reproducibility and intra- and inter-observer variability of tests.

We found that in situations that deviate only marginally from the classical diagnostic accuracy paradigm, for example, where there are few missing values on an otherwise acceptable reference standard or where the magnitude and type of imperfection in a reference standard is well documented, the methods we summarised may be valuable. However, in situations where an acceptable reference standard does not exist, holding on to the accuracy paradigm may be fruitless. In these situations, applying the concept of clinical test validation can provide a significant methodological advance (see also the next section of this chapter).

A different class of test evaluation methods focuses on changes in patient outcome from using tests. This class includes randomised clinical trials of tests: comparisons of testing versus not testing, or of one index test compared with another. As the number of testing strategies usually exceeds the number that can be adequately dealt with in an RCT, many researchers turn to modelling to explore the health consequences of testing. In decision analysis, data on test characteristics are combined with estimates of the effectiveness, risks and side-effects of treatment. A test's accuracy is not carved in stone, but will depend on the target condition that is to be detected. In talking about a test's accuracy, we will assume that the corresponding target condition has been defined.

If knowing a test's accuracy is important, the next question to ask is whether there is a reference standard providing adequate classification. This would be a reference standard about which there is consensus that it is the best available method for establishing the presence or absence of the target condition. In addition, researchers and readers should have confidence in the classification provided by reference standard, although misclassification should be considered in every accuracy study.

If the outcome of the reference standard cannot or has not been obtained in all study participants, statistical methods can be used to correct for the incomplete verification. The STARD statement encourages researchers to be explicit about missing information.[115] With small rates of missing information on reference standard outcome, these methods can be effectively used. For large volumes of missing data on the reference standard, the soundness of the approach depends on whether the mechanisms that led to the missing data are known. A general problem for these methods is that the general level of acceptance for correction techniques and for imputation methods is still fairly low.

If the pattern of missingness is not random, researchers could consider using a second

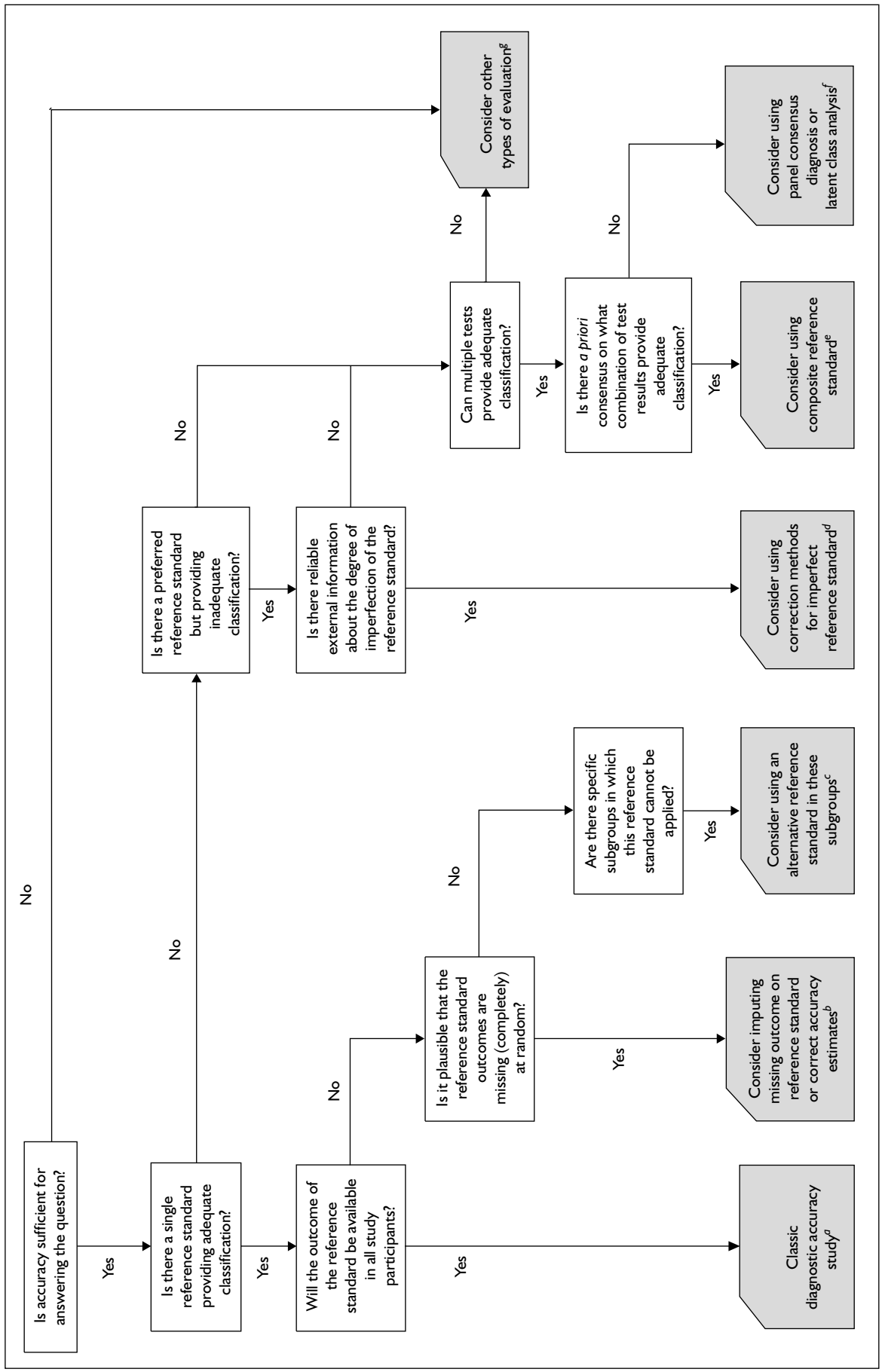


FIGURE 7 *Guidance for researchers when faced with no gold standard diagnostic research situations.*[a] *See Chapter 1.* [b] *See 'Impute or adjust for missing data on reference standard' (p. 13).* [c] *See 'Differential verification' (p.18).* [d] *See 'Correct imperfect reference standard' (p. 16).* [e] *See 'Composite reference standard' (p. 21).* [f] *See 'Panel or consensus diagnosis' (p. 24) and 'Latent data analysis' (p. 26).* [g] *See Chapter 5.*

reference standard in patients in whom the result of the first, intended reference standard is not available. The use of that second reference standard should preferably be stipulated in the protocol and reported as such. With differential verification, as in the use of one reference standard in test positive patients and a second one in test negative patients, the results should be reported for each reference standard to prevent bias.

If there is a reference standard but one that is known to be prone to error, statistical methods could be used to obtain estimates that are corrected for these reference standard errors. When there is an identified target condition but no available reference standard, the researcher may aim to construct one, based on two or more other tests. If these are used in a rule-based way to obtain information about the target condition, a composite reference standard has been constructed. If such a rule cannot be made explicit, it is possible to use a panel method, by inviting experts. An alternative method is the use of statistical techniques to obtain the most likely classification. By selecting one of these methods, the researchers – or the panel members – use their knowledge of the clinical target condition to identify variables or features that allow the identification of patients with that particular target condition.

With these techniques, there is an implicit assumption that they can be used to classify patients in previously unknown categories: the latent classes. This means that in latent class analysis the target condition is not defined in a clinical sense, but is a mathematically defined entity. The classification that comes out of the analysis may not fully coincide with pre-existing knowledge of the patients and the conditions that are amenable to treatment.

If none of these methods seem appropriate, the researchers have to turn to alternative methods for evaluation. Several proposals for a staged evaluation of tests have been made in the literature. A number of these focus on the added value or diagnostic gain of tests.[116] This includes studies that determine the added discriminatory power of tests, relative to pre-existing data from history, physical or previous tests. Multivariable logistic regression modelling is often used for this purpose, with the change in area under the curve as one of the statistics. The subjective side of the 'diagnostic gain' invites clinicians to express the reduction in experience they feel after having received the index test result. These reductions are elicited using visual analogue or quantitative scales. An additional step in these subjective approaches is to look at changes in management, either explicitly or in intended decisions.

Several authors have claimed that the accuracy paradigm is insufficient for evaluating the clinical usefulness of tests. The methods just mentioned can all be used, although most require an appropriate reference standard (diagnostic gain) or a clear link with management consequences (RCTs, decision analysis). We feel that there is both a need and a place for additional methods for exploring the likely clinical usefulness of tests. In the final section, below, we introduce the concept of test validation, a progressive exploration and testing of the tests results with other features, as a more general and more productive way of evaluating tests.

Limits to accuracy: towards a validation paradigm

In clinical medicine, tests are not only ordered for knowing the results for their own sake. A physician at the emergency department (ED) not only wants to know the concentration of cardiac troponin of the patient with chest pain, but also wants to know the likelihood of a diagnosis of acute coronary syndrome. In addition, he or she wants to know where to locate the patient in the risk spectrum of acute coronary syndrome.[77,117] The ED physician will probably rely on the troponin results and also on other findings in trying to find out what to do with the patient. Can this patient be safely sent home? Does the patient have to be admitted to the coronary care unit? Does the patient qualify for primary angioplasty? An umbrella question for these issues is: are patients with chest pain at the ED better off if they are routinely tested for troponin? In the answers to these questions, the exact value of the troponin concentration is an essential but insufficient element. It is not the attribute that is measured (troponin), but the level of confidence by which that attribute can be used to classify the patient in the risk spectrum of the acute coronary syndrome.

For decades, the diagnostic accuracy paradigm has been used to obtain answers to this type of question. It is difficult to pinpoint its origins exactly, but they go back well to – and beyond – Ledley and Lusted's seminal 1959 paper.[18] In its most classical sense, as can be found in all textbooks, the accuracy paradigm requires a gold standard, a technique used to identify with certainty patients with the disease of interest.

Pathophysiology and histology present the prototypical form of a gold standard: a method to determine unequivocally the presence or absence of disease in the human body. Sensitivity and specificity are the two best known statistics to express the results of the comparison of the test results with those of the gold standard.

The accuracy paradigm has proven to be a very valuable one. Its ubiquity has provided a focal point for the clinical evaluation of tests. The need for knowing the diagnostic accuracy of a test has prompted many useful evaluations. Reports of the poor sensitivity or specificity of a test are likely to inhibit the premature dissemination of tests in modern medicine.

The limited time frame of this report did not allow us to explore, analyse and speculate why the gold standard and the diagnostic accuracy paradigm have become so dominant in the evaluation of medical tests. We can only hint at the prevailing power of the classical view of clinical medicine, and its classical triad aetiognosis–diagnosis–prognosis. One can also point to the appealing simplicity of the basic accuracy design, which allows study results to be summarised in a very simple two-by-two table, or a slanted curve in two-dimensional ROC space. However, things should only be made as simple as possible and no simpler. Beyond any doubt, the prevailing accuracy paradigm is far from sufficient to cover all issues in the evaluation of the clinical evaluation of tests. We will summarise only a few of these issues here.

To start, many tests are not used for making a diagnosis at all. They are used for a variety of other purposes, such as guiding treatment decisions, monitoring treatment in chronic patients, informing patients and documenting changes in their condition, or for clinical or basic research. Many problems in present-day healthcare do not rely on issues of diagnosis, definitely not in chronic conditions, where patients are known to have diabetes, cardiovascular disease or COPD. The diagnostic accuracy paradigm cannot be simply applied in situations where tests are used for purposes other than diagnosis.

The second issue is the problematic relation between pathophysiology and diagnosis, and that between pathology and subsequent actions. Furthermore, the definition or the concept of a disease can change over time as new insights become available. The classical diagnosis of 'acute myocardial infarction' is now embedded in the spectrum of conditions known as 'acute coronary syndrome'. New management options and advances in testing can change the concept of disease, as in ectopic pregnancy. This diagnosis, once attached to a life-threatening condition, now covers subclinical conditions, such as 'throphoblast in regression'.[118] Advances in imaging allow the detection of even smaller, subsegmental pulmonary emboli or micro-metastases in patients with cancer. The current 'diagnoses' do not coincide with the older categories.

To some extent, this problem can be remedied by replacing the term 'disease' with 'target condition', as has been done in this report and elsewhere.[5] The target condition is a much broader concept, covering any particular disease, a disease stage or just any other identifiable condition that may prompt clinical action, such as further diagnostic testing or the initiation, modification or termination of treatment.

Another – and more widely documented – problem is the absence of a true gold standard to classify with certainty and without errors patients as having the target condition or not. In more cases than many imagine, such a gold standard simply does not exist. It definitely does not exist for many of the most prevalent chronic conditions, including diabetes, migraine and cardiovascular disease. For these problems, the term 'reference standard' has been introduced. This term acknowledges the absence of a 'gold standard' and refers to the best available method for classifying patients as having the target condition.

The term reference standard may seem like a solution, but in accordance with the Law of Frankenstein – 'the monster you create is yours, forever' – it also introduces ineradicable subjectivity. What should the reference standard be for appendicitis, or for deep venous thrombosis? Is it an image, or is it follow-up? The two methods will not yield identical results, and whereas one may be closer to pathophysiology, the other may be closer to patient outcome, the penultimate criterion for decisions in healthcare. A number of authors have suggested that these problems can be circumvented by jumping to the cornerstone of evaluations of interventions: the RCT. As summarised earlier in the report, RCTs cannot act as a panacea.[105,106] To allow meaningful interpretations of their results, such trials presume that the links between test results and subsequent clinical actions have been well established. That may not always be the case.

In our view, a move towards a test validation paradigm is justified. This means that scientists and

practitioners examine, using a number of different methods, whether the results of an index test are meaningful in practice. Validation will always be a gradual process. It will involve the scientific and clinical community defining a threshold, a point in the validation process, where the information gathered would be considered sufficient to allow clinical use of the test with confidence.

Validating a test is a process through which scientists and practitioners can find out whether the results of a test are meaningful. The process of validation will come after initial evaluations of the basic properties of the test: its reliability, consistency, trueness. Once these hurdles have been overcome, we can try to find out how and to what extent the test results fit in our understanding of a patient's condition, its likely causes and course.

To validate a test, we will have to build a conceptual framework, defining how the test results relate to other features. These features may be derived from history, from physical examination, from other test results, from follow-up or from response to treatment.

The concept of test validation encompasses the traditional notion of diagnostic accuracy. If there is an accepted pathophysiological gold standard, one that can be used to detect the presence of absence of disease with certainty, we can validate a test by comparing its results with the findings of the gold standard (criterion validity). Test validation also allows the option for evaluating test results of tests that are not used, or not used exclusively, in making a diagnosis.

Test validation is probably the way to go in case there is a test that is proclaimed to be 'better than the existing reference standard'. In these cases also, we have to show that the differences between that test and the reference standard are 'meaningful', by demonstrating reliable associations with other findings and test results.

The results of a validation process cannot be captured in simple statistics. The associations with other variables can and must be expressed in a quantitative sense, but they will never be reduced to a simple pair of numbers, as a test's sensitivity and specificity.

Discovering or demonstrating that test results are meaningful does not justify the use of that test in daily practice. In the end, the use of tests has to be justified by demonstrating that it leads to better healthcare, by improving outcome, reducing costs, or both.

Validation offers a way out of the dilemma introduced by the multiple fixes that are needed in order to cling to the diagnostic accuracy paradigm in the absence of an accepted gold standard. Replacing the gold standard by the reference standard, and the pathophysiological detection of disease by the identification of the target condition is not enough. Like all humans, we value simplicity, and the alluring attraction of two simple statistics makes it hard to give up the notion of fixed test characteristics. In some cases these fixes are feasible and reasonable, as has been summarised in this report. In others, we may well abandon the accuracy paradigm and replace it by a new one: the concept of the validation of medical tests.

Recommendations for further research

Diagnostic research with partial verification is common, this includes research based on routinely collected clinical data. There is a need to increase awareness that naïve estimates of accuracy by leaving out unverified patients from the calculation can lead to serious bias.

There is a need to perform empirical and simulation studies that will determine the potential and limitations of multiple imputation methods to reduce the bias caused by missing data on the reference standard.

In setting up consensus-based diagnosis, there are many practical choices to be made. These include the number of experts, the way patient information is presented, and how to obtain a final classification. There is a need to perform more methodological studies to provide guidance in this area.

There is a need to design studies that compare the results based on a consensus procedure with the results of latent class models when multiple pieces of information will be combined to construct a reference standard. Also there is a need to examine approaches that combine both procedures.

There is a need to develop more elaborate schemes outwith the accuracy paradigm for the validation of tests targeted at diseases or conditions for which there is no acceptable reference standard.

Acknowledgements

We gratefully acknowledge the many useful comments and interesting discussions we had with the many experts we involved in this project. For a list of experts, see Appendix 2. Additionally, we gained input from discussions with Professor Paul Glasziou, Dr Tracy Roberts, Dr Chris Hyde [CH is a member of the Editorial board for *Health Technology Assessment* but was not involved in the editorial process for this report] and Dr Priscilla Harries. We thank them for their time and the efforts that they put into this project. This report has been shaped and improved by the comments of the experts, but the views expressed in this report are those of the authors and do not necessarily reflect the opinions of the experts.

Contribution of authors

Anna Rutjes (Research Fellow) was a co-applicant on the project grant application and worked on the development of the protocol, acted as a reviewer, and was involved in the drafting of methods and results. Johannes Reitsma (Senior Clinical Epidemiologist) was a co-applicant on the project grant application and carried out work on the development of the protocol, project management, supervision of review work, as well as drafting and editing of the final report. Arri Coomarasamy (Honorary Lecturer in Epidemiology) was a co-applicant on the project grant application and worked on the development of the protocol and on drafting the results. Khalid Khan (Professor of Obstetrics-Gynaecology and Clinical Epidemiology) was a main applicant on the project grant application and also carried out development of the protocol, project management, drafting of results and editing of the final report. Patrick Bossuyt (Professor and Chair of Department of Clinical Epidemiology) was a main applicant on the project grant application and worked on the development of the protocol, on drafting the results and contributed to discussion, as well as to editing the final report.

References

1. Lijmer JG, Mol BW, Heisterkamp S, *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6.
2. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;**140**:189–202.
3. Rutjes AW, Reitsma JB, Di NM, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;**174**:469–76.
4. Knottnerus JA, van Weel C. General introduction: evaluation of diagnostic procedures. In Knottnerus JA, editor. *The evidence base of clinical diagnosis*. London: BMJ Books; 2002. pp. 1–18.
5. Sackett DL, Haynes RB. The architecture of diagnostic research. In Knottnerus JA, editor. *The evidence base of clinical diagnosis*. London: BMJ Books; 2002. pp. 19–38.
6. Lijmer JG, Mol BWJ, Bonsel GJ, Prins MH, Bossuyt PM. Strategies for the evaluation of diagnostic technologies. Evaluation of diagnostic tests: from accuracy to outcome (Thesis). Amsterdam; 2001. pp. 15–28.
7. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991;**11**: 88–94.
8. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In Knottnerus JA, editor. *The evidence base of clinical diagnosis*. London: BMJ Books; 2002. pp. 39–59.
9. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;**332**: 1089–92.
10. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med* 2006;**144**:850–5.
11. Boyko EJ, Alderman BW, Baron AE. Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. *J Gen Intern Med* 1988; **3**:476–81.
12. Deneef P. Evaluating rapid tests for streptococcal pharyngitis: the apparent accuracy of a diagnostic test when there are errors in the standard of comparison. *Med Decis Making* 1987;**7**:92–6.
13. Pepe MS. Incomplete data and imperfect reference tests. In Pepe MS, editor. *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press; 2004. pp. 168–213.
14. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* 1985;**41**:959–68.
15. Thibodeau L. Evaluating diagnostic tests. *Biometrics* 1981;**37**:801–4.
16. International Organization for Standardization. *International vocabulary of basic and general terms in metrology*. Geneva: ISO; 1993.
17. Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Rep* 1947;**62**:1432–49.
18. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science* 1959;**130**:9–21.
19. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;**140**:189–202.
20. Rutjes AW, Reitsma JB, Irwig L, Bossuyt PM. Partial and differential bias in diagnostic accuracy studies. Sources of bias and variation in diagnostic accuracy studies (Thesis). Amsterdam; 2005. pp. 31–44.
21. Gustafson P. The utility of prior information and stratification for parameter estimation with two screening tests but no gold standard. *Stat Med* 2005;**24**:1203–17.
22. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983;**39**:207–15.
23. Harel O, Zhou XH. Multiple imputation for correcting verification bias. *Stat Med* 2006;**25**: 3769–86.
24. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press; 2003.
25. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med* 1999;**18**:2987–3003.
26. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. *Stat Methods Med Res* 1998; **7**:354–70.

27. *Cochrane Handbook for Systematic Reviews of Interventions 4.2.5* [updated May 2005]. Chichester Wiley; 2006.

28. Lilford RJ, Richardson A, Stevens A, *et al*. Issues in methodological research: perspectives from researchers and commissioners. *Health Technol Assess* 2001;**5**(8).

29. Little R, Rubin D. *Statistical analysis with missing data*. 2nd ed. New York: Wiley; 2002.

30. Paliwal P, Gelfand AE. Estimating measures of diagnostic accuracy when some covariate information is missing. *Stat Med* 2006;**25**:2981–93.

31. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol* 1992;**45**:581–6.

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

33. Cecil MP, Kosinski AS, Jones MT, *et al*. The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. *J Clin Epidemiol* 1996;**49**:735–42.

34. Danias PG, Parker JA. Novel Internet-based tool for correcting apparent sensitivity and specificity of diagnostic tests to adjust for referral (verification) bias. *Radiographics* 2002;**22**(2):e4.

35. Held L, Ranyimbo AO. A Bayesian approach to estimate and validate the false negative fraction in a two-stage multiple screening test. *Methods Inf Med* 2004;**43**:461–4.

36. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics* 2003;**59**:163–71.

37. Kosinski AS, Barnhart HX. A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. *Stat Med* 2003;**22**:2711–21.

38. Toledano AY, Gatsonis C. Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics* 1999;**55**:488–96.

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

40. Zhou XH, Higgs RE. COMPROC and CHECKNORM: computer programs for comparing accuracies of diagnostic tests using ROC curves in the presence of verification bias. *Comput Methods Programs Biomed* 1998;**57**:179–86.

41. Zhou XH, Higgs RE. Assessing the relative accuracies of two screening tests in the presence of verification bias. *Stat Med* 2000;**19**:1697–705.

42. Molenberghs G, Goetghebeur EJ, Lipsitz SR, Kenward MG, Lesaffre E, Michiels B. Missing data perspectives of the fluvoxamine data set: a review. *Stat Med* 1999;**18**:2449–64.

43. Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. *J Clin Epidemiol* 2003;**56**:968–76.

44. Liu M, Taylor JM, Belin TR. Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics* 2000;**56**: 1157–63.

45. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc* 1996;**91**:473–89.

46. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res* 1999;**8**:3–15.

47. Zhou X-H, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine*. New York: Wiley; 2002.

48. Greenes RA, Begg CB. Assessment of diagnostic technologies. Methodology for unbiased estimation from samples of selectively verified patients. *Invest Radiol* 1985;**20**:751–6.

49. Pisano ED, Gatsonis C, Hendrick E, *et al*. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med* 2005;**353**:1773–83.

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

51. Drum D, Christacopoulos J. Hepatic scintigraphy in clinical decision making. *J Nucl Med* 1969; **13**:908–15.

52. Marshall V, Williams DC, Smith KD. Diaphanography as a means of detecting breast cancer. *Radiology* 1981;**150**:339–43.

53. Zhou X-H, Obuchowski NA, McClish DK. *Methods for correcting imperfect standard bias. Statistical methods in diagnostic medicine*. New York: Wiley; 2002. pp. 359–95.

54. Hadgu A, Dendukuri N, Hilden J. Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. *Epidemiology* 2005;**16**:604–12.

55. Staquet M, Rozencweig M, Lee YJ, *et al*. Methodology for the assessment of new dichotomous diagnostic tests. *J Chronic Dis* 1981;**34**:599–610.

56. Valenstein PN. Evaluating diagnostic tests with imperfect standards. *Am J Clin Pathol* 1990;**93**: 252–8.

57. Brenner H. Correcting for exposure misclassification using an alloyed gold standard. *Epidemiology* 1996;**7**:406–10.

58. Wacholder S, Armstrong B, Hartge P. Validation studies using an alloyed gold standard. *Am J Epidemiol* 1993;**137**:1251–8.

59. Schneeweiss S, Kriegmair M, Stepp H. Is everything all right if nothing seems wrong? A simple method of assessing the diagnostic value of endoscopic procedures when a gold standard is absent. *J Urol* 1999;**161**:1116–19.

60. Schneeweiss S. Sensitivity analysis of the diagnostic value of endoscopies in cross-sectional studies in the absence of a gold standard. *Int J Technol Assess Health Care* 2000;**16**:834–41.

61. van Belle A., Buller HR, Huisman MV, *et al*. Effectiveness of managing suspected pulmonary embolism using an algorithm combining clinical probability, D-dimer testing, and computed tomography. *JAMA* 2006;**295**:172–9.

62. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926–30.

63. Kline JA, Israel EG, Michelson EA, O'Neil BJ, Plewa MC, Portelli DC. Diagnostic accuracy of a bedside D-dimer assay and alveolar dead-space measurement for rapid exclusion of pulmonary embolism: a multicenter study. *JAMA* 2001;**285**:761–8.

64. Alonzo TA, Pepe MS. Assessing the accuracy of a new diagnostic test when a gold standard does not exist. *UW Biostatistics Working Paper Series* 1998; paper 156.

65. Hadgu A. The discrepancy in discrepant analysis. *Lancet* 1996;**348**:592–3.

66. Miller WC. Can we do better than discrepant analysis for new diagnostic test evaluation? *Clin Infect Dis* 1998;**27**:1186–93.

67. Miller WC. Bias in discrepant analysis: when two wrongs don't make a right. *J Clin Epidemiol* 1998; **51**:219–31.

68. Hadgu A. Bias in the evaluation of DNA-amplification tests for detecting *Chlamydia trachomatis*. *Stat Med* 1997;**16**:1391–9.

69. Lipman H, Astles J. Quantifying the bias associated with use of discrepant analysis. *Clin Chem* 1998;**44**:108–15.

70. Hadgu A. Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity. *J Clin Epidemiol* 1999;**52**:1231–7.

71. Hadgu A. Discrepant analysis is an inappropriate and unscientific method. *J Clin Microbiol* 2000;**38**:4301–2.

72. Green TA, Black CM, Johnson RE. Evaluation of bias in diagnostic-test sensitivity and specificity estimates computed by discrepant analysis. *J Clin Microbiol* 1998;**36**:375–81.

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

74. Wu CH, Lee MF, Yin SC, Yang DM, Cheng SF. Comparison of polymerase chain reaction, monoclonal antibody based enzyme immunoassay, and cell culture for detection of *Chlamydia trachomatis* in genital specimens. *Sex Transm Dis* 1992;**19**:193–7.

75. Martin DH, Nsuami M, Schachter J, *et al*. Use of multiple nucleic acid amplification tests to define the infected-patient "gold standard" in clinical trials of new diagnostic tests for *Chlamydia trachomatis* infections. *J Clin Microbiol* 2004; **42**:4749–58.

76. Jang D, Sellors JW, Mahony JB, Pickard L, Chernesky MA. Effects of broadening the gold standard on the performance of a chemiluminometric immunoassay to detect *Chlamydia trachomatis* antigens in centrifuged first void urine and urethral swab samples from men. *Sex Transm Dis* 1992;**19**:315–19.

77. Das R, Kilcullen N, Morrell C, Robinson MB, Barth JH, Hall AS. The British Cardiac Society Working Group definition of myocardial infarction: implications for practice. *Heart* 2006;**92**:21–6.

78. Thijs JC, Van Zwet AA, Thijs WJ, *et al*. Diagnostic tests for *Helicobacter pylori*: a prospective evaluation of their accuracy, without selecting a single test as the gold standard. *Am J Gastroenterol* 1996; **91**:2125–9.

79. Madan K, Ahuja V, Gupta SD, Bal C, Kapoor A, Sharma MP. Impact of 24-h esophageal pH monitoring on the diagnosis of gastroesophageal reflux disease: defining the gold standard. *J Gastroenterol Hepatol* 2005;**20**:30–7.

80. Gagnon R, Charlin B, Coletti M, Sauve E, Van D, V. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Med Educ* 2005;**39**:284–91.

81. Jones J, Hunter D. Consensus methods for medical and health services research. *BMJ* 1995;**311**:376–80.

82. Rutten FH, Moons KG, Cramer MJ, *et al*. Recognising heart failure in elderly patients with stable chronic obstructive pulmonary disease in primary care: cross sectional diagnostic study. *BMJ* 2005;**331**:1379.

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

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

85. Pepe MS, Janes H. Insights into latent class analysis. *UW Biostatistics Working Paper Series* 2005; paper 236.

86. Formann AK. Measurement errors in caries diagnosis: some further latent class models. *Biometrics* 1994;**50**:865–71.

87. Committee on Quality Management, Japan Society of Clinical Chemistry. Proposed standard for matrix reference materials for quantitative laboratory methods. *JPN J Clin Chem* 2003; **32**:180–5.

88. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* 1996;**52**:797–810.

89. Bertrand P, Benichou J, Grenier P, Chastang C. Hui and Walter's latent-class reference-free approach may be more useful in assessing agreement than diagnostic performance. *J Clin Epidemiol* 2005;**58**:688–700.

90. Vermunt JK, Magidson J. Latent class analysis. In Lewis-Beck M, Bryman A, Liao TF, editors. *The Sage Encyclopedia of Social Sciences Research Methods*. Thousand Oaks, CA Sage Publications; 2004. pp. 549–53.

91. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol* 1995;**141**:263–72.

92. Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. *Stat Med* 2002;**21**:2653–69.

93. Hadgu A, Qu Y. A biomedical application of latent class models with random effects. *Appl Stat* 1998;**47**:603–16.

94. Gustafson P. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Boca Raton, FL: Chapman and Hall/CRC; 2003.

95. Dendukuri N, Joseph L. Bayesian approaches to modelling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001; **57**:158–67.

96. Georgiadis MP, Johnson WO, Gardner IA, Singh R. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Appl Stat* 2003;**52**:63–76.

97. Bertrand P, Benichou J, Grenier P, Chastang C. Hui and Walter's latent-class reference-free approach may be more useful in assessing agreement than diagnostic performance. *J Clin Epidemiol* 2005;**58**:688–700.

98. Black CM. Current methods of laboratory diagnosis of *Chlamydia trachomatis* infections. *Clin Microbiol Rev* 1997;**10**:160–84.

99. Barnes RC. Laboratory diagnosis of human chlamydial infections. *Clin Microbiol Rev* 1989;**2**:119–36.

100. Winter G. *A comparative discussion of the notion of 'validity' in qualitative and quantitative research. The qualitative report*. 2000; 4. URL: http://www.nova.edu/ssss/QR/OR4-3/winter.html. Accessed 21 September 2007.

101. Streiner DL, Norman GR. Validity. In Streiner DL, Norman GR, editors. *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press; 1995. pp. 144–62.

102. Bland JM, Altman DG. Statistics notes: validating scales and indexes. *BMJ* 2002;**324**:606–7.

103. Fletcher RH, Fletcher SW, Wagner EH. Abnormality. In Satterfield TS, editor. *Clinical epidemiology: the essentials*. Baltimore, MD: Williams and Wilkins; 1996. pp. 22–24.

104. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;**52**:281–302.

105. Lijmer JG, Bossuyt PM. Diagnostic testing and prognosis: the randomised controlled trial in diagnostic research. In Knottnerus JA, editor. *The evidence base of clinical diagnosis*. London: BMJ Books; 2002. pp. 61–80.

106. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000;**356**:1844–7.

107. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991;**11**: 88–94.

108. Subtil D, Goeusse P, Houfflin-Debarge V, *et al*. Randomised comparison of uterine artery Doppler and aspirin (100 mg) with placebo in nulliparous women: the Essai Regional Aspirine Mere-Enfant Study (Part 2). *BJOG* 2003;**110**:485–91.

109. Duley L, Henderson-Smart D, Knight M, King J. Antiplatelet drugs for prevention of pre-eclampsia and its consequences: systematic review. *BMJ* 2001;**322**:329–33.

110. Collinson PO, Stubbs PJ, Kessler AC. Multicentre evaluation of the diagnostic value of cardiac troponin T, CK-MB mass, and myoglobin for assessing patients with suspected acute coronary syndromes in routine clinical practice. *Heart* 2003;**89**:280–6.

111. Kontos MC, Shah R, Fritz LM, *et al*. Implication of different cardiac troponin I levels for clinical outcomes and prognosis of acute chest pain patients. *J Am Coll Cardiol* 2004;**43**:958–65.

112. Jaffe AS, Ravkilde J, Roberts R, *et al*. It's time for a change to a troponin standard. *Circulation* 2000;**102**:1216–20.

113. Apple FS, Wu AH, Mair J, *et al*. Future biomarkers for detection of ischemia and risk stratification in acute coronary syndrome. *Clin Chem* 2005;**51**:810–24.

114. Ewer K, Deeks J, Alvarez L, *et al*. Comparison of T-cell-based assay with tuberculin skin test for diagnosis of Mycobacterium tuberculosis infection in a school tuberculosis outbreak. *Lancet* 2003;**361**:1168–73.

115. Bossuyt PM, Reitsma JB, Bruns DE, *et al*. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003;**49**:7–18.

116. Moons KG, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. *Clin Chem* 2004;**50**:473–6.

117. Lee TH, Goldman L. Evaluation of the patient with acute chest pain. *N Engl J Med* 2000; **342**:1187–95.

118. Ankum WM, Van der Veen F, Hamerlynck JV, Lammes FB. Suspected ectopic pregnancy. What to do when human chorionic gonadotropin levels are below the discriminatory zone. *J Reprod Med* 1995;**40**:525–8.

Appendix 1

Search terms in databases

MEDLINE
Searched 6 October 2005 via OVID
Date coverage: 1966–September 2005
1. reference.ti.
2. gold.ti.
3. golden.ti.
4. test.ti.
5. standard.ti.
6. 1 or 2 or 3
7. 4 or 5
8. 6 and 7
9. absen$.ti.
10. reference test.ti.
11. 9 and 10
12. 9 and 1
13. 9 and 2
14. 9 and 3
15. 10 and 5
16. 8 or 11 or 12 or 13 or 14 or 15

Cochrane databases
(DARE, CENTRAL, CMR, NHS)
Searched 13 October 2005
1. reference.ti.
2. gold.ti.
3. golden.ti.
4. test.ti.
5. standard.ti.
6. 1 or 2 or 3
7. 4 or 5
8. 6 and 7
9. absen*.ti.
10. reference test.ti.
11. 9 and 10
12. 10 and 1
13. 9 and 2
14 9 and 3
15 9 and 5
16 8 or 11 or 12 or 13 or 14 or 15

Medion
Searched 24 November 2005 via www.mediondatabase.nl
Filter used: 'Methodological Studies on Systematic Reviews of Diagnostic Tests'
1. MKD (quality assessment)
2. MBD (bias in meta-analysis)
3. MH (heterogeneity)
4. MAD (several/methods)

EMBASE
Searched 6 October 2005 via OVID
Date coverage: 1980–September 2005
1. reference.ti.
2. gold.ti.
3. golden.ti.
4. test.ti.
5. standard.ti.
6. 1 or 2 or 3
7. 4 or 5
8. 6 and 7
9. absen$.ti.
10. reference test.ti.
11. 9 and 11
12. 9 and 1
13. 9 and 2
14. 9 and 3
15. 10 and 5
16. 8 or 11 or 12 or 13 or 14 or 15

Pubmed
Searched 6 October 2005 via www.pubmed.gov
Date coverage: 1966–October 2005
((("reference"[ti] OR "gold"[ti] OR "golden"[ti]) AND("test"[ti] OR "standard"[ti])) OR ((absen*[ti] AND gold*[ti]) OR ("no gold*"[ti]) OR (absen*[ti] AND referen*[ti]) OR (absen*[ti] AND standard[ti]) OR (absen*[ti] AND "reference test"[ti])))

Appendix 2

Experts in peer review process

We assembled a list of topic-specific experts outside our research team to review each method. In addition, several general experts reviewed the whole report.

Topic-specific experts

'Impute missing data on reference standard' and 'Correct imperfect reference standard'

Professor Aeilko H Zwinderman
Professor in Biostatistics, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, The Netherlands

Dr Francisca Galindo Garre
Statistician, Psychometrician, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, The Netherlands

'Construct reference standard'

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

Professor Karel GM Moons
Professor of Clinical Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht, The Netherlands

Dr Alexandra AH van Abswoude
Methodologist, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, The Netherlands

'Validate index test results'

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

General experts reviewing the full report

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

Professor Karel GM Moons
Professor of Clinical Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht, The Netherlands

Professor Aeilko H. Zwinderman
Professor of Biostatistics, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, The Netherlands

Professor Constantine Gatsonis
Professor of Biostatistics, Community Health (Biostatistics) and Applied Mathematics Brown University, Providence, RI, USA

Health Technology Assessment reports published to date

Volume 1, 1997

No. 1
Home parenteral nutrition: a systematic review.
By Richards DM, Deeks JJ, Sheldon TA, Shaffer JL.

No. 2
Diagnosis, management and screening of early localised prostate cancer.
A review by Selley S, Donovan J, Faulkner A, Coast J, Gillatt D.

No. 3
The diagnosis, management, treatment and costs of prostate cancer in England and Wales.
A review by Chamberlain J, Melia J, Moss S, Brown J.

No. 4
Screening for fragile X syndrome.
A review by Murray J, Cuckle H, Taylor G, Hewison J.

No. 5
A review of near patient testing in primary care.
By Hobbs FDR, Delaney BC, Fitzmaurice DA, Wilson S, Hyde CJ, Thorpe GH, *et al*.

No. 6
Systematic review of outpatient services for chronic pain control.
By McQuay HJ, Moore RA, Eccleston C, Morley S, de C Williams AC.

No. 7
Neonatal screening for inborn errors of metabolism: cost, yield and outcome.
A review by Pollitt RJ, Green A, McCabe CJ, Booth A, Cooper NJ, Leonard JV, *et al*.

No. 8
Preschool vision screening.
A review by Snowdon SK, Stewart-Brown SL.

No. 9
Implications of socio-cultural contexts for the ethics of clinical trials.
A review by Ashcroft RE, Chadwick DW, Clark SRL, Edwards RHT, Frith L, Hutton JL.

No. 10
A critical review of the role of neonatal hearing screening in the detection of congenital hearing impairment.
By Davis A, Bamford J, Wilson I, Ramkalawan T, Forshaw M, Wright S.

No. 11
Newborn screening for inborn errors of metabolism: a systematic review.
By Seymour CA, Thomason MJ, Chalmers RA, Addison GM, Bain MD, Cockburn F, *et al*.

No. 12
Routine preoperative testing: a systematic review of the evidence.
By Munro J, Booth A, Nicholl J.

No. 13
Systematic review of the effectiveness of laxatives in the elderly.
By Petticrew M, Watt I, Sheldon T.

No. 14
When and how to assess fast-changing technologies: a comparative study of medical applications of four generic technologies.
A review by Mowatt G, Bower DJ, Brebner JA, Cairns JA, Grant AM, McKee L.

Volume 2, 1998

No. 1
Antenatal screening for Down's syndrome.
A review by Wald NJ, Kennard A, Hackshaw A, McGuire A.

No. 2
Screening for ovarian cancer: a systematic review.
By Bell R, Petticrew M, Luengo S, Sheldon TA.

No. 3
Consensus development methods, and their use in clinical guideline development.
A review by Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CFB, Askham J, *et al*.

No. 4
A cost–utility analysis of interferon beta for multiple sclerosis.
By Parkin D, McNamee P, Jacoby A, Miller P, Thomas S, Bates D.

No. 5
Effectiveness and efficiency of methods of dialysis therapy for end-stage renal disease: systematic reviews.
By MacLeod A, Grant A, Donaldson C, Khan I, Campbell M, Daly C, *et al*.

No. 6
Effectiveness of hip prostheses in primary total hip replacement: a critical review of evidence and an economic model.
By Faulkner A, Kennedy LG, Baxter K, Donovan J, Wilkinson M, Bevan G.

No. 7
Antimicrobial prophylaxis in colorectal surgery: a systematic review of randomised controlled trials.
By Song F, Glenny AM.

No. 8
Bone marrow and peripheral blood stem cell transplantation for malignancy.
A review by Johnson PWM, Simnett SJ, Sweetenham JW, Morgan GJ, Stewart LA.

No. 9
Screening for speech and language delay: a systematic review of the literature.
By Law J, Boyle J, Harris F, Harkness A, Nye C.

No. 10
Resource allocation for chronic stable angina: a systematic review of effectiveness, costs and cost-effectiveness of alternative interventions.
By Sculpher MJ, Petticrew M, Kelland JL, Elliott RA, Holdright DR, Buxton MJ.

No. 11
Detection, adherence and control of hypertension for the prevention of stroke: a systematic review.
By Ebrahim S.

No. 12
Postoperative analgesia and vomiting, with special reference to day-case surgery: a systematic review.
By McQuay HJ, Moore RA.

No. 13
Choosing between randomised and nonrandomised studies: a systematic review.
By Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C.

No. 14
Evaluating patient-based outcome measures for use in clinical trials.
A review by Fitzpatrick R, Davey C, Buxton MJ, Jones DR.

No. 15
Ethical issues in the design and conduct of randomised controlled trials.
A review by Edwards SJL, Lilford RJ, Braunholtz DA, Jackson JC, Hewison J, Thornton J.

No. 16
Qualitative research methods in health technology assessment: a review of the literature.
By Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P.

No. 17
The costs and benefits of paramedic skills in pre-hospital trauma care.
By Nicholl J, Hughes S, Dixon S, Turner J, Yates D.

No. 18
Systematic review of endoscopic ultrasound in gastro-oesophageal cancer.
By Harris KM, Kelly S, Berry E, Hutton J, Roderick P, Cullingworth J, *et al*.

No. 19
Systematic reviews of trials and other studies.
By Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F.

No. 20
Primary total hip replacement surgery: a systematic review of outcomes and modelling of cost-effectiveness associated with different prostheses.
A review by Fitzpatrick R, Shortall E, Sculpher M, Murray D, Morris R, Lodge M, *et al*.

Volume 3, 1999

No. 1
Informed decision making: an annotated bibliography and systematic review.
By Bekker H, Thornton JG, Airey CM, Connelly JB, Hewison J, Robinson MB, *et al*.

No. 2
Handling uncertainty when performing economic evaluation of healthcare interventions.
A review by Briggs AH, Gray AM.

No. 3
The role of expectancies in the placebo effect and their use in the delivery of health care: a systematic review.
By Crow R, Gage H, Hampson S, Hart J, Kimber A, Thomas H.

No. 4
A randomised controlled trial of different approaches to universal antenatal HIV testing: uptake and acceptability. Annex: Antenatal HIV testing – assessment of a routine voluntary approach.
By Simpson WM, Johnstone FD, Boyd FM, Goldberg DJ, Hart GJ, Gormley SM, *et al*.

No. 5
Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review.
By Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ.

No. 6
Assessing the costs of healthcare technologies in clinical trials.
A review by Johnston K, Buxton MJ, Jones DR, Fitzpatrick R.

No. 7
Cooperatives and their primary care emergency centres: organisation and impact.
By Hallam L, Henthorne K.

No. 8
Screening for cystic fibrosis.
A review by Murray J, Cuckle H, Taylor G, Littlewood J, Hewison J.

No. 9
A review of the use of health status measures in economic evaluation.
By Brazier J, Deverill M, Green C, Harper R, Booth A.

No. 10
Methods for the analysis of quality-of-life and survival data in health technology assessment.
A review by Billingham LJ, Abrams KR, Jones DR.

No. 11
Antenatal and neonatal haemoglobinopathy screening in the UK: review and economic analysis.
By Zeuner D, Ades AE, Karnon J, Brown J, Dezateux C, Anionwu EN.

No. 12
Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses.
A review by Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, *et al*.

No. 13
'Early warning systems' for identifying new healthcare technologies.
By Robert G, Stevens A, Gabbay J.

No. 14
A systematic review of the role of human papillomavirus testing within a cervical screening programme.
By Cuzick J, Sasieni P, Davies P, Adams J, Normand C, Frater A, *et al*.

No. 15
Near patient testing in diabetes clinics: appraising the costs and outcomes.
By Grieve R, Beech R, Vincent J, Mazurkiewicz J.

No. 16
Positron emission tomography: establishing priorities for health technology assessment.
A review by Robert G, Milne R.

No. 17 (Pt 1)
The debridement of chronic wounds: a systematic review.
By Bradley M, Cullum N, Sheldon T.

No. 17 (Pt 2)
Systematic reviews of wound care management: (2) Dressings and topical agents used in the healing of chronic wounds.
By Bradley M, Cullum N, Nelson EA, Petticrew M, Sheldon T, Torgerson D.

No. 18
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.
By Berry E, Kelly S, Hutton J, Harris KM, Roderick P, Boyce JC, *et al*.

No. 19
What role for statins? A review and economic model.
By Ebrahim S, Davey Smith G, McCabe C, Payne N, Pickin M, Sheldon TA, *et al*.

No. 20
Factors that limit the quality, number and progress of randomised controlled trials.
A review by Prescott RJ, Counsell CE, Gillespie WJ, Grant AM, Russell IT, Kiauka S, *et al*.

No. 21
Antimicrobial prophylaxis in total hip replacement: a systematic review.
By Glenny AM, Song F.

No. 22
Health promoting schools and health promotion in schools: two systematic reviews.
By Lister-Sharp D, Chapman S, Stewart-Brown S, Sowden A.

No. 23
Economic evaluation of a primary care-based education programme for patients with osteoarthritis of the knee.
A review by Lord J, Victor C, Littlejohns P, Ross FM, Axford JS.

Volume 4, 2000

No. 1
The estimation of marginal time preference in a UK-wide sample (TEMPUS) project.
A review by Cairns JA, van der Pol MM.

No. 2
Geriatric rehabilitation following fractures in older people: a systematic review.
By Cameron I, Crotty M, Currie C, Finnegan T, Gillespie L, Gillespie W, *et al*.

No. 3
Screening for sickle cell disease and thalassaemia: a systematic review with supplementary research.
By Davies SC, Cronin E, Gill M, Greengross P, Hickman M, Normand C.

No. 4
Community provision of hearing aids and related audiology services.
A review by Reeves DJ, Alborz A, Hickson FS, Bamford JM.

No. 5
False-negative results in screening programmes: systematic review of impact and implications.
By Petticrew MP, Sowden AJ, Lister-Sharp D, Wright K.

No. 6
Costs and benefits of community postnatal support workers: a randomised controlled trial.
By Morrell CJ, Spiby H, Stewart P, Walters S, Morgan A.

No. 7
Implantable contraceptives (subdermal implants and hormonally impregnated intrauterine systems) versus other forms of reversible contraceptives: two systematic reviews to assess relative effectiveness, acceptability, tolerability and cost-effectiveness.
By French RS, Cowan FM, Mansour DJA, Morris S, Procter T, Hughes D, *et al*.

No. 8
An introduction to statistical methods for health technology assessment.
A review by White SJ, Ashby D, Brown PJ.

No. 9
Disease-modifying drugs for multiple sclerosis: a rapid and systematic review.
By Clegg A, Bryant J, Milne R.

No. 10
Publication and related biases.
A review by Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ.

No. 11
Cost and outcome implications of the organisation of vascular services.
By Michaels J, Brazier J, Palfreyman S, Shackley P, Slack R.

No. 12
Monitoring blood glucose control in diabetes mellitus: a systematic review.
By Coster S, Gulliford MC, Seed PT, Powrie JK, Swaminathan R.

No. 13
The effectiveness of domiciliary health visiting: a systematic review of international studies and a selective review of the British literature.
By Elkan R, Kendrick D, Hewitt M, Robinson JJA, Tolley K, Blair M, *et al*.

No. 14
The determinants of screening uptake and interventions for increasing uptake: a systematic review.
By Jepson R, Clegg A, Forbes C, Lewis R, Sowden A, Kleijnen J.

No. 15
The effectiveness and cost-effectiveness of prophylactic removal of wisdom teeth.
A rapid review by Song F, O'Meara S, Wilson P, Golder S, Kleijnen J.

No. 16
Ultrasound screening in pregnancy: a systematic review of the clinical effectiveness, cost-effectiveness and women's views.
By Bricker L, Garcia J, Henderson J, Mugford M, Neilson J, Roberts T, *et al*.

No. 17
A rapid and systematic review of the effectiveness and cost-effectiveness of the taxanes used in the treatment of advanced breast and ovarian cancer.
By Lister-Sharp D, McDonagh MS, Khan KS, Kleijnen J.

No. 18
Liquid-based cytology in cervical screening: a rapid and systematic review.
By Payne N, Chilcott J, McGoogan E.

No. 19
Randomised controlled trial of non-directive counselling, cognitive–behaviour therapy and usual general practitioner care in the management of depression as well as mixed anxiety and depression in primary care.
By King M, Sibbald B, Ward E, Bower P, Lloyd M, Gabbay M, *et al*.

No. 20
Routine referral for radiography of patients presenting with low back pain: is patients' outcome influenced by GPs' referral for plain radiography?
By Kerry S, Hilton S, Patel S, Dundas D, Rink E, Lord J.

No. 21
Systematic reviews of wound care management: (3) antimicrobial agents for chronic wounds; (4) diabetic foot ulceration.
By O'Meara S, Cullum N, Majid M, Sheldon T.

No. 22
Using routine data to complement and enhance the results of randomised controlled trials.
By Lewsey JD, Leyland AH, Murray GD, Boddy FA.

No. 23
Coronary artery stents in the treatment of ischaemic heart disease: a rapid and systematic review.
By Meads C, Cummins C, Jolly K, Stevens A, Burls A, Hyde C.

No. 24
Outcome measures for adult critical care: a systematic review.
By Hayes JA, Black NA, Jenkinson C, Young JD, Rowan KM, Daly K, *et al*.

No. 25
A systematic review to evaluate the effectiveness of interventions to promote the initiation of breastfeeding.
By Fairbank L, O'Meara S, Renfrew MJ, Woolridge M, Sowden AJ, Lister-Sharp D.

No. 26
Implantable cardioverter defibrillators: arrhythmias. A rapid and systematic review.
By Parkes J, Bryant J, Milne R.

No. 27
Treatments for fatigue in multiple sclerosis: a rapid and systematic review.
By Brañas P, Jordan R, Fry-Smith A, Burls A, Hyde C.

No. 28
Early asthma prophylaxis, natural history, skeletal development and economy (EASE): a pilot randomised controlled trial.
By Baxter-Jones ADG, Helms PJ, Russell G, Grant A, Ross S, Cairns JA, *et al*.

No. 29
Screening for hypercholesterolaemia versus case finding for familial hypercholesterolaemia: a systematic review and cost-effectiveness analysis.
By Marks D, Wonderling D, Thorogood M, Lambert H, Humphries SE, Neil HAW.

No. 30
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of glycoprotein IIb/IIIa antagonists in the medical management of unstable angina.
By McDonagh MS, Bachmann LM, Golder S, Kleijnen J, ter Riet G.

No. 31
A randomised controlled trial of prehospital intravenous fluid replacement therapy in serious trauma.
By Turner J, Nicholl J, Webber L, Cox H, Dixon S, Yates D.

No. 32
Intrathecal pumps for giving opioids in chronic pain: a systematic review.
By Williams JE, Louw G, Towlerton G.

No. 33
Combination therapy (interferon alfa and ribavirin) in the treatment of chronic hepatitis C: a rapid and systematic review.
By Shepherd J, Waugh N, Hewitson P.

No. 34
A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies.
By MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AMS.

No. 35
Intravascular ultrasound-guided interventions in coronary artery disease: a systematic literature review, with decision-analytic modelling, of outcomes and cost-effectiveness.
By Berry E, Kelly S, Hutton J, Lindsay HSJ, Blaxill JM, Evans JA, *et al*.

No. 36
A randomised controlled trial to evaluate the effectiveness and cost-effectiveness of counselling patients with chronic depression.
By Simpson S, Corney R, Fitzgerald P, Beecham J.

No. 37
Systematic review of treatments for atopic eczema.
By Hoare C, Li Wan Po A, Williams H.

No. 38
Bayesian methods in health technology assessment: a review.
By Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR.

No. 39
The management of dyspepsia: a systematic review.
By Delaney B, Moayyedi P, Deeks J, Innes M, Soo S, Barton P, *et al*.

No. 40
A systematic review of treatments for severe psoriasis.
By Griffiths CEM, Clark CM, Chalmers RJG, Li Wan Po A, Williams HC.

Volume 5, 2001

No. 1
Clinical and cost-effectiveness of donepezil, rivastigmine and galantamine for Alzheimer's disease: a rapid and systematic review.
By Clegg A, Bryant J, Nicholson T, McIntyre L, De Broe S, Gerard K, *et al*.

No. 2
The clinical effectiveness and cost-effectiveness of riluzole for motor neurone disease: a rapid and systematic review.
By Stewart A, Sandercock J, Bryan S, Hyde C, Barton PM, Fry-Smith A, *et al*.

No. 3
Equity and the economic evaluation of healthcare.
By Sassi F, Archard L, Le Grand J.

No. 4
Quality-of-life measures in chronic diseases of childhood.
By Eiser C, Morse R.

No. 5
Eliciting public preferences for healthcare: a systematic review of techniques.
By Ryan M, Scott DA, Reeves C, Bate A, van Teijlingen ER, Russell EM, *et al*.

No. 6
General health status measures for people with cognitive impairment: learning disability and acquired brain injury.
By Riemsma RP, Forbes CA, Glanville JM, Eastwood AJ, Kleijnen J.

No. 7
An assessment of screening strategies for fragile X syndrome in the UK.
By Pembrey ME, Barnicoat AJ, Carmichael B, Bobrow M, Turner G.

No. 8
Issues in methodological research: perspectives from researchers and commissioners.
By Lilford RJ, Richardson A, Stevens A, Fitzpatrick R, Edwards S, Rock F, *et al*.

No. 9
Systematic reviews of wound care management: (5) beds; (6) compression; (7) laser therapy, therapeutic ultrasound, electrotherapy and electromagnetic therapy.
By Cullum N, Nelson EA, Flemming K, Sheldon T.

No. 10
Effects of educational and psychosocial interventions for adolescents with diabetes mellitus: a systematic review.
By Hampson SE, Skinner TC, Hart J, Storey L, Gage H, Foxcroft D, *et al*.

No. 11
Effectiveness of autologous chondrocyte transplantation for hyaline cartilage defects in knees: a rapid and systematic review.
By Jobanputra P, Parry D, Fry-Smith A, Burls A.

No. 12
Statistical assessment of the learning curves of health technologies.
By Ramsay CR, Grant AM, Wallace SA, Garthwaite PH, Monk AF, Russell IT.

No. 13
The effectiveness and cost-effectiveness of temozolomide for the treatment of recurrent malignant glioma: a rapid and systematic review.
By Dinnes J, Cave C, Huang S, Major K, Milne R.

No. 14
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of debriding agents in treating surgical wounds healing by secondary intention.
By Lewis R, Whiting P, ter Riet G, O'Meara S, Glanville J.

No. 15
Home treatment for mental health problems: a systematic review.
By Burns T, Knapp M, Catty J, Healey A, Henderson J, Watt H, *et al*.

No. 16
How to develop cost-conscious guidelines.
By Eccles M, Mason J.

No. 17
The role of specialist nurses in multiple sclerosis: a rapid and systematic review.
By De Broe S, Christopher F, Waugh N.

No. 18
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of orlistat in the management of obesity.
By O'Meara S, Riemsma R, Shirran L, Mather L, ter Riet G.

No. 19
The clinical effectiveness and cost-effectiveness of pioglitazone for type 2 diabetes mellitus: a rapid and systematic review.
By Chilcott J, Wight J, Lloyd Jones M, Tappenden P.

No. 20
Extended scope of nursing practice: a multicentre randomised controlled trial of appropriately trained nurses and preregistration house officers in pre-operative assessment in elective general surgery.
By Kinley H, Czoski-Murray C, George S, McCabe C, Primrose J, Reilly C, *et al*.

No. 21
Systematic reviews of the effectiveness of day care for people with severe mental disorders: (1) Acute day hospital versus admission; (2) Vocational rehabilitation; (3) Day hospital versus outpatient care.
By Marshall M, Crowther R, Almaraz-Serrano A, Creed F, Sledge W, Kluiter H, *et al*.

No. 22
The measurement and monitoring of surgical adverse events.
By Bruce J, Russell EM, Mollison J, Krukowski ZH.

No. 23
Action research: a systematic review and guidance for assessment.
By Waterman H, Tillen D, Dickson R, de Koning K.

No. 24
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of gemcitabine for the treatment of pancreatic cancer.
By Ward S, Morris E, Bansback N, Calvert N, Crellin A, Forman D, *et al*.

No. 25
A rapid and systematic review of the evidence for the clinical effectiveness and cost-effectiveness of irinotecan, oxaliplatin and raltitrexed for the treatment of advanced colorectal cancer.
By Lloyd Jones M, Hummel S, Bansback N, Orr B, Seymour M.

No. 26
Comparison of the effectiveness of inhaler devices in asthma and chronic obstructive airways disease: a systematic review of the literature.
By Brocklebank D, Ram F, Wright J, Barry P, Cates C, Davies L, *et al*.

No. 27
The cost-effectiveness of magnetic resonance imaging for investigation of the knee joint.
By Bryan S, Weatherburn G, Bungay H, Hatrick C, Salas C, Parry D, *et al*.

No. 28
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of topotecan for ovarian cancer.
By Forbes C, Shirran L, Bagnall A-M, Duffy S, ter Riet G.

No. 29
Superseded by a report published in a later volume.

No. 30
The role of radiography in primary care patients with low back pain of at least 6 weeks duration: a randomised (unblinded) controlled trial.
By Kendrick D, Fielding K, Bentley E, Miller P, Kerslake R, Pringle M.

No. 31
Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients.
By McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, *et al*.

No. 32
A rapid and systematic review of the clinical effectiveness and cost-effectiveness of paclitaxel, docetaxel, gemcitabine and vinorelbine in non-small-cell lung cancer.
By Clegg A, Scott DA, Sidhu M, Hewitson P, Waugh N.

No. 33
Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives.
By Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G.

No. 34
Depot antipsychotic medication in the treatment of patients with schizophrenia: (1) Meta-review; (2) Patient and nurse attitudes.
By David AS, Adams C.

No. 35
A systematic review of controlled trials of the effectiveness and cost-effectiveness of brief psychological treatments for depression.
By Churchill R, Hunot V, Corney R, Knapp M, McGuire H, Tylee A, *et al*.

No. 36
Cost analysis of child health surveillance.
By Sanderson D, Wright D, Acton C, Duree D.

Volume 6, 2002

No. 1
A study of the methods used to select review criteria for clinical audit.
By Hearnshaw H, Harker R, Cheater F, Baker R, Grimshaw G.

No. 2
Fludarabine as second-line therapy for B cell chronic lymphocytic leukaemia: a technology assessment.
By Hyde C, Wake B, Bryan S, Barton P, Fry-Smith A, Davenport C, *et al*.

No. 3
Rituximab as third-line treatment for refractory or recurrent Stage III or IV follicular non-Hodgkin's lymphoma: a systematic review and economic evaluation.
By Wake B, Hyde C, Bryan S, Barton P, Song F, Fry-Smith A, *et al*.

No. 4
A systematic review of discharge arrangements for older people.
By Parker SG, Peet SM, McPherson A, Cannaby AM, Baker R, Wilson A, *et al*.

No. 5
The clinical effectiveness and cost-effectiveness of inhaler devices used in the routine management of chronic asthma in older children: a systematic review and economic evaluation.
By Peters J, Stevenson M, Beverley C, Lim J, Smith S.

No. 6
The clinical effectiveness and cost-effectiveness of sibutramine in the management of obesity: a technology assessment.
By O'Meara S, Riemsma R, Shirran L, Mather L, ter Riet G.

No. 7
The cost-effectiveness of magnetic resonance angiography for carotid artery stenosis and peripheral vascular disease: a systematic review.
By Berry E, Kelly S, Westwood ME, Davies LM, Gough MJ, Bamford JM, *et al*.

No. 8
Promoting physical activity in South Asian Muslim women through 'exercise on prescription'.
By Carroll B, Ali N, Azam N.

No. 9
Zanamivir for the treatment of influenza in adults: a systematic review and economic evaluation.
By Burls A, Clark W, Stewart T, Preston C, Bryan S, Jefferson T, *et al*.

No. 10
A review of the natural history and epidemiology of multiple sclerosis: implications for resource allocation and health economic models.
By Richards RG, Sampson FC, Beard SM, Tappenden P.

No. 11
Screening for gestational diabetes: a systematic review and economic evaluation.
By Scott DA, Loveman E, McIntyre L, Waugh N.

No. 12
The clinical effectiveness and cost-effectiveness of surgery for people with morbid obesity: a systematic review and economic evaluation.
By Clegg AJ, Colquitt J, Sidhu MK, Royle P, Loveman E, Walker A.

No. 13
The clinical effectiveness of trastuzumab for breast cancer: a systematic review.
By Lewis R, Bagnall A-M, Forbes C, Shirran E, Duffy S, Kleijnen J, *et al*.

No. 14
The clinical effectiveness and cost-effectiveness of vinorelbine for breast cancer: a systematic review and economic evaluation.
By Lewis R, Bagnall A-M, King S, Woolacott N, Forbes C, Shirran L, *et al*.

No. 15
A systematic review of the effectiveness and cost-effectiveness of metal-on-metal hip resurfacing arthroplasty for treatment of hip disease.
By Vale L, Wyness L, McCormack K, McKenzie L, Brazzelli M, Stearns SC.

No. 16
The clinical effectiveness and cost-effectiveness of bupropion and nicotine replacement therapy for smoking cessation: a systematic review and economic evaluation.
By Woolacott NF, Jones L, Forbes CA, Mather LC, Sowden AJ, Song FJ, *et al*.

No. 17
A systematic review of effectiveness and economic evaluation of new drug treatments for juvenile idiopathic arthritis: etanercept.
By Cummins C, Connock M, Fry-Smith A, Burls A.

No. 18
Clinical effectiveness and cost-effectiveness of growth hormone in children: a systematic review and economic evaluation.
By Bryant J, Cave C, Mihaylova B, Chase D, McIntyre L, Gerard K, *et al*.

No. 19
Clinical effectiveness and cost-effectiveness of growth hormone in adults in relation to impact on quality of life: a systematic review and economic evaluation.
By Bryant J, Loveman E, Chase D, Mihaylova B, Cave C, Gerard K, *et al*.

No. 20
Clinical medication review by a pharmacist of patients on repeat prescriptions in general practice: a randomised controlled trial.
By Zermansky AG, Petty DR, Raynor DK, Lowe CJ, Freementle N, Vail A.

No. 21
The effectiveness of infliximab and etanercept for the treatment of rheumatoid arthritis: a systematic review and economic evaluation.
By Jobanputra P, Barton P, Bryan S, Burls A.

No. 22
A systematic review and economic evaluation of computerised cognitive behaviour therapy for depression and anxiety.
By Kaltenthaler E, Shackley P, Stevens K, Beverley C, Parry G, Chilcott J.

No. 23
A systematic review and economic evaluation of pegylated liposomal doxorubicin hydrochloride for ovarian cancer.
By Forbes C, Wilby J, Richardson G, Sculpher M, Mather L, Reimsma R.

No. 24
A systematic review of the effectiveness of interventions based on a stages-of-change approach to promote individual behaviour change.
By Riemsma RP, Pattenden J, Bridle C, Sowden AJ, Mather L, Watt IS, *et al*.

No. 25
A systematic review update of the clinical effectiveness and cost-effectiveness of glycoprotein IIb/IIIa antagonists.
By Robinson M, Ginnelly L, Sculpher M, Jones L, Riemsma R, Palmer S, *et al*.

No. 26
A systematic review of the effectiveness, cost-effectiveness and barriers to implementation of thrombolytic and neuroprotective therapy for acute ischaemic stroke in the NHS.
By Sandercock P, Berge E, Dennis M, Forbes J, Hand P, Kwan J, *et al*.

No. 27
A randomised controlled crossover trial of nurse practitioner versus doctor-led outpatient care in a bronchiectasis clinic.
By Caine N, Sharples LD, Hollingworth W, French J, Keogan M, Exley A, *et al*.

No. 28
Clinical effectiveness and cost – consequences of selective serotonin reuptake inhibitors in the treatment of sex offenders.
By Adi Y, Ashcroft D, Browne K, Beech A, Fry-Smith A, Hyde C.

No. 29
Treatment of established osteoporosis: a systematic review and cost–utility analysis.
By Kanis JA, Brazier JE, Stevenson M, Calvert NW, Lloyd Jones M.

No. 30
Which anaesthetic agents are cost-effective in day surgery? Literature review, national survey of practice and randomised controlled trial.
By Elliott RA Payne K, Moore JK, Davies LM, Harper NJN, St Leger AS, *et al*.

No. 31
Screening for hepatitis C among injecting drug users and in genitourinary medicine clinics: systematic reviews of effectiveness, modelling study and national survey of current practice.
By Stein K, Dalziel K, Walker A, McIntyre L, Jenkins B, Horne J, *et al*.

No. 32
The measurement of satisfaction with healthcare: implications for practice from a systematic review of the literature.
By Crow R, Gage H, Hampson S, Hart J, Kimber A, Storey L, *et al*.

No. 33
The effectiveness and cost-effectiveness of imatinib in chronic myeloid leukaemia: a systematic review.
By Garside R, Round A, Dalziel K, Stein K, Royle R.

No. 34
A comparative study of hypertonic saline, daily and alternate-day rhDNase in children with cystic fibrosis.
By Suri R, Wallis C, Bush A, Thompson S, Normand C, Flather M, *et al*.

No. 35
A systematic review of the costs and effectiveness of different models of paediatric home care.
By Parker G, Bhakta P, Lovett CA, Paisley S, Olsen R, Turner D, *et al*.

Volume 7, 2003

No. 1
How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study.
By Egger M, Jüni P, Bartlett C, Holenstein F, Sterne J.

No. 2
Systematic review of the effectiveness and cost-effectiveness, and economic evaluation, of home versus hospital or satellite unit haemodialysis for people with end-stage renal failure.
By Mowatt G, Vale L, Perez J, Wyness L, Fraser C, MacLeod A, *et al*.

No. 3
Systematic review and economic evaluation of the effectiveness of infliximab for the treatment of Crohn's disease.
By Clark W, Raftery J, Barton P, Song F, Fry-Smith A, Burls A.

No. 4
A review of the clinical effectiveness and cost-effectiveness of routine anti-D prophylaxis for pregnant women who are rhesus negative.
By Chilcott J, Lloyd Jones M, Wight J, Forman K, Wray J, Beverley C, *et al*.

No. 5
Systematic review and evaluation of the use of tumour markers in paediatric oncology: Ewing's sarcoma and neuroblastoma.
By Riley RD, Burchill SA, Abrams KR, Heney D, Lambert PC, Jones DR, *et al*.

No. 6
The cost-effectiveness of screening for *Helicobacter pylori* to reduce mortality and morbidity from gastric cancer and peptic ulcer disease: a discrete-event simulation model.
By Roderick P, Davies R, Raftery J, Crabbe D, Pearce R, Bhandari P, *et al*.

No. 7
The clinical effectiveness and cost-effectiveness of routine dental checks: a systematic review and economic evaluation.
By Davenport C, Elley K, Salas C, Taylor-Weetman CL, Fry-Smith A, Bryan S, *et al*.

No. 8
A multicentre randomised controlled trial assessing the costs and benefits of using structured information and analysis of women's preferences in the management of menorrhagia.
By Kennedy ADM, Sculpher MJ, Coulter A, Dwyer N, Rees M, Horsley S, *et al*.

No. 9
Clinical effectiveness and cost–utility of photodynamic therapy for wet age-related macular degeneration: a systematic review and economic evaluation.
By Meads C, Salas C, Roberts T, Moore D, Fry-Smith A, Hyde C.

No. 10
Evaluation of molecular tests for prenatal diagnosis of chromosome abnormalities.
By Grimshaw GM, Szczepura A, Hultén M, MacDonald F, Nevin NC, Sutton F, *et al*.

No. 11
First and second trimester antenatal screening for Down's syndrome: the results of the Serum, Urine and Ultrasound Screening Study (SURUSS).
By Wald NJ, Rodeck C, Hackshaw AK, Walters J, Chitty L, Mackinson AM.

No. 12
The effectiveness and cost-effectiveness of ultrasound locating devices for central venous access: a systematic review and economic evaluation.
By Calvert N, Hind D, McWilliams RG, Thomas SM, Beverley C, Davidson A.

No. 13
A systematic review of atypical antipsychotics in schizophrenia.
By Bagnall A-M, Jones L, Lewis R, Ginnelly L, Glanville J, Torgerson D, *et al*.

No. 14
Prostate Testing for Cancer and Treatment (ProtecT) feasibility study.
By Donovan J, Hamdy F, Neal D, Peters T, Oliver S, Brindle L, *et al*.

No. 15
Early thrombolysis for the treatment of acute myocardial infarction: a systematic review and economic evaluation.
By Boland A, Dundar Y, Bagust A, Haycox A, Hill R, Mujica Mota R, *et al*.

No. 16
Screening for fragile X syndrome: a literature review and modelling.
By Song FJ, Barton P, Sleightholme V, Yao GL, Fry-Smith A.

No. 17
Systematic review of endoscopic sinus surgery for nasal polyps.
By Dalziel K, Stein K, Round A, Garside R, Royle P.

No. 18
Towards efficient guidelines: how to monitor guideline use in primary care.
By Hutchinson A, McIntosh A, Cox S, Gilbert C.

No. 19
Effectiveness and cost-effectiveness of acute hospital-based spinal cord injuries services: systematic review.
By Bagnall A-M, Jones L, Richardson G, Duffy S, Riemsma R.

No. 20
Prioritisation of health technology assessment. The PATHS model: methods and case studies.
By Townsend J, Buxton M, Harper G.

No. 21
Systematic review of the clinical effectiveness and cost-effectiveness of tension-free vaginal tape for treatment of urinary stress incontinence.
By Cody J, Wyness L, Wallace S, Glazener C, Kilonzo M, Stearns S, *et al*.

No. 22
The clinical and cost-effectiveness of patient education models for diabetes: a systematic review and economic evaluation.
By Loveman E, Cave C, Green C, Royle P, Dunn N, Waugh N.

No. 23
The role of modelling in prioritising and planning clinical trials.
By Chilcott J, Brennan A, Booth A, Karnon J, Tappenden P.

No. 24
Cost–benefit evaluation of routine influenza immunisation in people 65–74 years of age.
By Allsup S, Gosney M, Haycox A, Regan M.

No. 25
The clinical and cost-effectiveness of pulsatile machine perfusion versus cold storage of kidneys for transplantation retrieved from heart-beating and non-heart-beating donors.
By Wight J, Chilcott J, Holmes M, Brewer N.

No. 26
Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment.
By Williams JG, Cheung WY, Cohen DR, Hutchings HA, Longo MF, Russell IT.

No. 27
Evaluating non-randomised intervention studies.
By Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, *et al*.

No. 28
A randomised controlled trial to assess the impact of a package comprising a patient-orientated, evidence-based self-help guidebook and patient-centred consultations on disease management and satisfaction in inflammatory bowel disease.
By Kennedy A, Nelson E, Reeves D, Richardson G, Roberts C, Robinson A, *et al*.

No. 29
The effectiveness of diagnostic tests for the assessment of shoulder pain due to soft tissue disorders: a systematic review.
By Dinnes J, Loveman E, McIntyre L, Waugh N.

No. 30
The value of digital imaging in diabetic retinopathy.
By Sharp PF, Olson J, Strachan F, Hipwell J, Ludbrook A, O'Donnell M, *et al*.

No. 31
Lowering blood pressure to prevent myocardial infarction and stroke: a new preventive strategy.
By Law M, Wald N, Morris J.

No. 32
Clinical and cost-effectiveness of capecitabine and tegafur with uracil for the treatment of metastatic colorectal cancer: systematic review and economic evaluation.
By Ward S, Kaltenthaler E, Cowan J, Brewer N.

No. 33
Clinical and cost-effectiveness of new and emerging technologies for early localised prostate cancer: a systematic review.
By Hummel S, Paisley S, Morgan A, Currie E, Brewer N.

No. 34
Literature searching for clinical and cost-effectiveness studies used in health technology assessment reports carried out for the National Institute for Clinical Excellence appraisal system.
By Royle P, Waugh N.

No. 35
Systematic review and economic decision modelling for the prevention and treatment of influenza A and B.
By Turner D, Wailoo A, Nicholson K, Cooper N, Sutton A, Abrams K.

No. 36
A randomised controlled trial to evaluate the clinical and cost-effectiveness of Hickman line insertions in adult cancer patients by nurses.
By Boland A, Haycox A, Bagust A, Fitzsimmons L.

No. 37
Redesigning postnatal care: a randomised controlled trial of protocol-based midwifery-led care focused on individual women's physical and psychological health needs.
By MacArthur C, Winter HR, Bick DE, Lilford RJ, Lancashire RJ, Knowles H, *et al*.

No. 38
Estimating implied rates of discount in healthcare decision-making.
By West RR, McNabb R, Thompson AGH, Sheldon TA, Grimley Evans J.

No. 39
Systematic review of isolation policies in the hospital management of methicillin-resistant *Staphylococcus aureus*: a review of the literature with epidemiological and economic modelling.
By Cooper BS, Stone SP, Kibbler CC, Cookson BD, Roberts JA, Medley GF, *et al*.

No. 40
Treatments for spasticity and pain in multiple sclerosis: a systematic review.
By Beard S, Hunn A, Wight J.

No. 41
The inclusion of reports of randomised trials published in languages other than English in systematic reviews.
By Moher D, Pham B, Lawson ML, Klassen TP.

No. 42
The impact of screening on future health-promoting behaviours and health beliefs: a systematic review.
By Bankhead CR, Brett J, Bukach C, Webster P, Stewart-Brown S, Munafo M, *et al*.

Volume 8, 2004

No. 1
What is the best imaging strategy for acute stroke?
By Wardlaw JM, Keir SL, Seymour J, Lewis S, Sandercock PAG, Dennis MS, *et al*.

No. 2
Systematic review and modelling of the investigation of acute and chronic chest pain presenting in primary care.
By Mant J, McManus RJ, Oakes RAL, Delaney BC, Barton PM, Deeks JJ, *et al*.

No. 3
The effectiveness and cost-effectiveness of microwave and thermal balloon endometrial ablation for heavy menstrual bleeding: a systematic review and economic modelling.
By Garside R, Stein K, Wyatt K, Round A, Price A.

No. 4
A systematic review of the role of bisphosphonates in metastatic disease.
By Ross JR, Saunders Y, Edmonds PM, Patel S, Wonderling D, Normand C, *et al*.

No. 5
Systematic review of the clinical effectiveness and cost-effectiveness of capecitabine (Xeloda®) for locally advanced and/or metastatic breast cancer.
By Jones L, Hawkins N, Westwood M, Wright K, Richardson G, Riemsma R.

No. 6
Effectiveness and efficiency of guideline dissemination and implementation strategies.
By Grimshaw JM, Thomas RE, MacLennan G, Fraser C, Ramsay CR, Vale L, *et al*.

No. 7
Clinical effectiveness and costs of the Sugarbaker procedure for the treatment of pseudomyxoma peritonei.
By Bryant J, Clegg AJ, Sidhu MK, Brodin H, Royle P, Davidson P.

No. 8
Psychological treatment for insomnia in the regulation of long-term hypnotic drug use.
By Morgan K, Dixon S, Mathers N, Thompson J, Tomeny M.

No. 9
Improving the evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome.
By Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ.

No. 10
A systematic review and economic evaluation of magnetic resonance cholangiopancreatography compared with diagnostic endoscopic retrograde cholangiopancreatography.
By Kaltenthaler E, Bravo Vergel Y, Chilcott J, Thomas S, Blakeborough T, Walters SJ, *et al*.

No. 11
The use of modelling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis.
By Barton P, Jobanputra P, Wilson J, Bryan S, Burls A.

No. 12
Clinical effectiveness and cost-effectiveness of neonatal screening for inborn errors of metabolism using tandem mass spectrometry: a systematic review.
By Pandor A, Eastham J, Beverley C, Chilcott J, Paisley S.

No. 13
Clinical effectiveness and cost-effectiveness of pioglitazone and rosiglitazone in the treatment of type 2 diabetes: a systematic review and economic evaluation.
By Czoski-Murray C, Warren E, Chilcott J, Beverley C, Psyllaki MA, Cowan J.

No. 14
Routine examination of the newborn: the EMREN study. Evaluation of an extension of the midwife role including a randomised controlled trial of appropriately trained midwives and paediatric senior house officers.
By Townsend J, Wolke D, Hayes J, Davé S, Rogers C, Bloomfield L, *et al*.

No. 15
Involving consumers in research and development agenda setting for the NHS: developing an evidence-based approach.
By Oliver S, Clarke-Jones L, Rees R, Milne R, Buchanan P, Gabbay J, *et al*.

No. 16
A multi-centre randomised controlled trial of minimally invasive direct coronary bypass grafting versus percutaneous transluminal coronary angioplasty with stenting for proximal stenosis of the left anterior descending coronary artery.
By Reeves BC, Angelini GD, Bryan AJ, Taylor FC, Cripps T, Spyt TJ, *et al*.

No. 17
Does early magnetic resonance imaging influence management or improve outcome in patients referred to secondary care with low back pain? A pragmatic randomised controlled trial.
By Gilbert FJ, Grant AM, Gillan MGC, Vale L, Scott NW, Campbell MK, *et al*.

No. 18
The clinical and cost-effectiveness of anakinra for the treatment of rheumatoid arthritis in adults: a systematic review and economic analysis.
By Clark W, Jobanputra P, Barton P, Burls A.

No. 19
A rapid and systematic review and economic evaluation of the clinical and cost-effectiveness of newer drugs for treatment of mania associated with bipolar affective disorder.
By Bridle C, Palmer S, Bagnall A-M, Darba J, Duffy S, Sculpher M, *et al*.

No. 20
Liquid-based cytology in cervical screening: an updated rapid and systematic review and economic analysis.
By Karnon J, Peters J, Platt J, Chilcott J, McGoogan E, Brewer N.

No. 21
Systematic review of the long-term effects and economic consequences of treatments for obesity and implications for health improvement.
By Avenell A, Broom J, Brown TJ, Poobalan A, Aucott L, Stearns SC, *et al*.

No. 22
Autoantibody testing in children with newly diagnosed type 1 diabetes mellitus.
By Dretzke J, Cummins C, Sandercock J, Fry-Smith A, Barrett T, Burls A.

No. 23
Clinical effectiveness and cost-effectiveness of prehospital intravenous fluids in trauma patients.
By Dretzke J, Sandercock J, Bayliss S, Burls A.

No. 24
Newer hypnotic drugs for the short-term management of insomnia: a systematic review and economic evaluation.
By Dündar Y, Boland A, Strobl J, Dodd S, Haycox A, Bagust A, *et al*.

No. 25
Development and validation of methods for assessing the quality of diagnostic accuracy studies.
By Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J.

No. 26
EVALUATE hysterectomy trial: a multicentre randomised trial comparing abdominal, vaginal and laparoscopic methods of hysterectomy.
By Garry R, Fountain J, Brown J, Manca A, Mason S, Sculpher M, *et al*.

No. 27
Methods for expected value of information analysis in complex health economic models: developments on the health economics of interferon-β and glatiramer acetate for multiple sclerosis.
By Tappenden P, Chilcott JB, Eggington S, Oakley J, McCabe C.

No. 28
Effectiveness and cost-effectiveness of imatinib for first-line treatment of chronic myeloid leukaemia in chronic phase: a systematic review and economic analysis.
By Dalziel K, Round A, Stein K, Garside R, Price A.

No. 29
VenUS I: a randomised controlled trial of two types of bandage for treating venous leg ulcers.
By Iglesias C, Nelson EA, Cullum NA, Torgerson DJ on behalf of the VenUS Team.

No. 30
Systematic review of the effectiveness and cost-effectiveness, and economic evaluation, of myocardial perfusion scintigraphy for the diagnosis and management of angina and myocardial infarction.
By Mowatt G, Vale L, Brazzelli M, Hernandez R, Murray A, Scott N, *et al*.

No. 31
A pilot study on the use of decision theory and value of information analysis as part of the NHS Health Technology Assessment programme.
By Claxton K, Ginnelly L, Sculpher M, Philips Z, Palmer S.

No. 32
The Social Support and Family Health Study: a randomised controlled trial and economic evaluation of two alternative forms of postnatal support for mothers living in disadvantaged inner-city areas.
By Wiggins M, Oakley A, Roberts I, Turner H, Rajan L, Austerberry H, *et al*.

No. 33
Psychosocial aspects of genetic screening of pregnant women and newborns: a systematic review.
By Green JM, Hewison J, Bekker HL, Bryant, Cuckle HS.

No. 34
Evaluation of abnormal uterine bleeding: comparison of three outpatient procedures within cohorts defined by age and menopausal status.
By Critchley HOD, Warner P, Lee AJ, Brechin S, Guise J, Graham B.

No. 35
Coronary artery stents: a rapid systematic review and economic evaluation.
By Hill R, Bagust A, Bakhai A, Dickson R, Dündar Y, Haycox A, *et al*.

No. 36
Review of guidelines for good practice in decision-analytic modelling in health technology assessment.
By Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, *et al*.

No. 37
Rituximab (MabThera®) for aggressive non-Hodgkin's lymphoma: systematic review and economic evaluation.
By Knight C, Hind D, Brewer N, Abbott V.

No. 38
Clinical effectiveness and cost-effectiveness of clopidogrel and modified-release dipyridamole in the secondary prevention of occlusive vascular events: a systematic review and economic evaluation.
By Jones L, Griffin S, Palmer S, Main C, Orton V, Sculpher M, *et al.*

No. 39
Pegylated interferon α-2a and -2b in combination with ribavirin in the treatment of chronic hepatitis C: a systematic review and economic evaluation.
By Shepherd J, Brodin H, Cave C, Waugh N, Price A, Gabbay J.

No. 40
Clopidogrel used in combination with aspirin compared with aspirin alone in the treatment of non-ST-segment-elevation acute coronary syndromes: a systematic review and economic evaluation.
By Main C, Palmer S, Griffin S, Jones L, Orton V, Sculpher M, *et al.*

No. 41
Provision, uptake and cost of cardiac rehabilitation programmes: improving services to under-represented groups.
By Beswick AD, Rees K, Griebsch I, Taylor FC, Burke M, West RR, *et al.*

No. 42
Involving South Asian patients in clinical trials.
By Hussain-Gambles M, Leese B, Atkin K, Brown J, Mason S, Tovey P.

No. 43
Clinical and cost-effectiveness of continuous subcutaneous insulin infusion for diabetes.
By Colquitt JL, Green C, Sidhu MK, Hartwell D, Waugh N.

No. 44
Identification and assessment of ongoing trials in health technology assessment reviews.
By Song FJ, Fry-Smith A, Davenport C, Bayliss S, Adi Y, Wilson JS, *et al*.

No. 45
Systematic review and economic evaluation of a long-acting insulin analogue, insulin glargine
By Warren E, Weatherley-Jones E, Chilcott J, Beverley C.

No. 46
Supplementation of a home-based exercise programme with a class-based programme for people with osteoarthritis of the knees: a randomised controlled trial and health economic analysis.
By McCarthy CJ, Mills PM, Pullen R, Richardson G, Hawkins N, Roberts CR, *et al*.

No. 47
Clinical and cost-effectiveness of once-daily versus more frequent use of same potency topical corticosteroids for atopic eczema: a systematic review and economic evaluation.
By Green C, Colquitt JL, Kirby J, Davidson P, Payne E.

No. 48
Acupuncture of chronic headache disorders in primary care: randomised controlled trial and economic analysis.
By Vickers AJ, Rees RW, Zollman CE, McCarney R, Smith CM, Ellis N, *et al*.

No. 49
Generalisability in economic evaluation studies in healthcare: a review and case studies.
By Sculpher MJ, Pang FS, Manca A, Drummond MF, Golder S, Urdahl H, *et al*.

No. 50
Virtual outreach: a randomised controlled trial and economic evaluation of joint teleconferenced medical consultations.
By Wallace P, Barber J, Clayton W, Currell R, Fleming K, Garner P, *et al*.

Volume 9, 2005

No. 1
Randomised controlled multiple treatment comparison to provide a cost-effectiveness rationale for the selection of antimicrobial therapy in acne.
By Ozolins M, Eady EA, Avery A, Cunliffe WJ, O'Neill C, Simpson NB, *et al*.

No. 2
Do the findings of case series studies vary significantly according to methodological characteristics?
By Dalziel K, Round A, Stein K, Garside R, Castelnuovo E, Payne L.

No. 3
Improving the referral process for familial breast cancer genetic counselling: findings of three randomised controlled trials of two interventions.
By Wilson BJ, Torrance N, Mollison J, Wordsworth S, Gray JR, Haites NE, *et al*.

No. 4
Randomised evaluation of alternative electrosurgical modalities to treat bladder outflow obstruction in men with benign prostatic hyperplasia.
By Fowler C, McAllister W, Plail R, Karim O, Yang Q.

No. 5
A pragmatic randomised controlled trial of the cost-effectiveness of palliative therapies for patients with inoperable oesophageal cancer.
By Shenfine J, McNamee P, Steen N, Bond J, Griffin SM.

No. 6
Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography.
By Taylor P, Champness J, Given-Wilson R, Johnston K, Potts H.

No. 7
Issues in data monitoring and interim analysis of trials.
By Grant AM, Altman DG, Babiker AB, Campbell MK, Clemens FJ, Darbyshire JH, *et al*.

No. 8
Lay public's understanding of equipoise and randomisation in randomised controlled trials.
By Robinson EJ, Kerr CEP, Stevens AJ, Lilford RJ, Braunholtz DA, Edwards SJ, *et al*.

No. 9
Clinical and cost-effectiveness of electroconvulsive therapy for depressive illness, schizophrenia, catatonia and mania: systematic reviews and economic modelling studies.
By Greenhalgh J, Knight C, Hind D, Beverley C, Walters S.

No. 10
Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology.
By Smith SC, Lamping DL, Banerjee S, Harwood R, Foley B, Smith P, *et al*.

No. 11
Clinical effectiveness and cost-effectiveness of drotrecogin alfa (activated) (Xigris®) for the treatment of severe sepsis in adults: a systematic review and economic evaluation.
By Green C, Dinnes J, Takeda A, Shepherd J, Hartwell D, Cave C, *et al*.

No. 12
A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy.
By Dinnes J, Deeks J, Kirby J, Roderick P.

No. 13
Cervical screening programmes: can automation help? Evidence from systematic reviews, an economic analysis and a simulation modelling exercise applied to the UK.
By Willis BH, Barton P, Pearmain P, Bryan S, Hyde C.

No. 14
Laparoscopic surgery for inguinal hernia repair: systematic review of effectiveness and economic evaluation.
By McCormack K, Wake B, Perez J, Fraser C, Cook J, McIntosh E, *et al*.

No. 15
Clinical effectiveness, tolerability and cost-effectiveness of newer drugs for epilepsy in adults: a systematic review and economic evaluation.
By Wilby J, Kainth A, Hawkins N, Epstein D, McIntosh H, McDaid C, *et al*.

No. 16
A randomised controlled trial to compare the cost-effectiveness of tricyclic antidepressants, selective serotonin reuptake inhibitors and lofepramine.
By Peveler R, Kendrick T, Buxton M, Longworth L, Baldwin D, Moore M, *et al*.

No. 17
Clinical effectiveness and cost-effectiveness of immediate angioplasty for acute myocardial infarction: systematic review and economic evaluation.
By Hartwell D, Colquitt J, Loveman E, Clegg AJ, Brodin H, Waugh N, *et al*.

No. 18
A randomised controlled comparison of alternative strategies in stroke care.
By Kalra L, Evans A, Perez I, Knapp M, Swift C, Donaldson N.

No. 19
The investigation and analysis of critical incidents and adverse events in healthcare.
By Woloshynowych M, Rogers S, Taylor-Adams S, Vincent C.

No. 20
Potential use of routine databases in health technology assessment.
By Raftery J, Roderick P, Stevens A.

No. 21
Clinical and cost-effectiveness of newer immunosuppressive regimens in renal transplantation: a systematic review and modelling study.
By Woodroffe R, Yao GL, Meads C, Bayliss S, Ready A, Raftery J, *et al*.

No. 22
A systematic review and economic evaluation of alendronate, etidronate, risedronate, raloxifene and teriparatide for the prevention and treatment of postmenopausal osteoporosis.
By Stevenson M, Lloyd Jones M, De Nigris E, Brewer N, Davis S, Oakley J.

No. 23
A systematic review to examine the impact of psycho-educational interventions on health outcomes and costs in adults and children with difficult asthma.
By Smith JR, Mugford M, Holland R, Candy B, Noble MJ, Harrison BDW, *et al*.

No. 24
An evaluation of the costs, effectiveness and quality of renal replacement therapy provision in renal satellite units in England and Wales.
By Roderick P, Nicholson T, Armitage A, Mehta R, Mullee M, Gerard K, *et al*.

No. 25
Imatinib for the treatment of patients with unresectable and/or metastatic gastrointestinal stromal tumours: systematic review and economic evaluation.
By Wilson J, Connock M, Song F, Yao G, Fry-Smith A, Raftery J, *et al*.

No. 26
Indirect comparisons of competing interventions.
By Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, *et al*.

No. 27
Cost-effectiveness of alternative strategies for the initial medical management of non-ST elevation acute coronary syndrome: systematic review and decision-analytical modelling.
By Robinson M, Palmer S, Sculpher M, Philips Z, Ginnelly L, Bowens A, *et al*.

No. 28
Outcomes of electrically stimulated gracilis neosphincter surgery.
By Tillin T, Chambers M, Feldman R.

No. 29
The effectiveness and cost-effectiveness of pimecrolimus and tacrolimus for atopic eczema: a systematic review and economic evaluation.
By Garside R, Stein K, Castelnuovo E, Pitt M, Ashcroft D, Dimmock P, *et al*.

No. 30
Systematic review on urine albumin testing for early detection of diabetic complications.
By Newman DJ, Mattock MB, Dawnay ABS, Kerry S, McGuire A, Yaqoob M, *et al*.

No. 31
Randomised controlled trial of the cost-effectiveness of water-based therapy for lower limb osteoarthritis.
By Cochrane T, Davey RC, Matthes Edwards SM.

No. 32
Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain.
By Thomas KJ, MacPherson H, Ratcliffe J, Thorpe L, Brazier J, Campbell M, *et al*.

No. 33
Cost-effectiveness and safety of epidural steroids in the management of sciatica.
By Price C, Arden N, Coglan L, Rogers P.

No. 34
The British Rheumatoid Outcome Study Group (BROSG) randomised controlled trial to compare the effectiveness and cost-effectiveness of aggressive versus symptomatic therapy in established rheumatoid arthritis.
By Symmons D, Tricker K, Roberts C, Davies L, Dawes P, Scott DL.

No. 35
Conceptual framework and systematic review of the effects of participants' and professionals' preferences in randomised controlled trials.
By King M, Nazareth I, Lampe F, Bower P, Chandler M, Morou M, *et al*.

No. 36
The clinical and cost-effectiveness of implantable cardioverter defibrillators: a systematic review.
By Bryant J, Brodin H, Loveman E, Payne E, Clegg A.

No. 37
A trial of problem-solving by community mental health nurses for anxiety, depression and life difficulties among general practice patients. The CPN-GP study.
By Kendrick T, Simons L, Mynors-Wallis L, Gray A, Lathlean J, Pickering R, *et al*.

No. 38
The causes and effects of socio-demographic exclusions from clinical trials.
By Bartlett C, Doyal L, Ebrahim S, Davey P, Bachmann M, Egger M, *et al*.

No. 39
Is hydrotherapy cost-effective? A randomised controlled trial of combined hydrotherapy programmes compared with physiotherapy land techniques in children with juvenile idiopathic arthritis.
By Epps H, Ginnelly L, Utley M, Southwood T, Gallivan S, Sculpher M, *et al*.

No. 40
A randomised controlled trial and cost-effectiveness study of systematic screening (targeted and total population screening) versus routine practice for the detection of atrial fibrillation in people aged 65 and over. The SAFE study.
By Hobbs FDR, Fitzmaurice DA, Mant J, Murray E, Jowett S, Bryan S, *et al*.

No. 41
Displaced intracapsular hip fractures in fit, older people: a randomised comparison of reduction and fixation, bipolar hemiarthroplasty and total hip arthroplasty.
By Keating JF, Grant A, Masson M, Scott NW, Forbes JF.

No. 42
Long-term outcome of cognitive behaviour therapy clinical trials in central Scotland.
By Durham RC, Chambers JA, Power KG, Sharp DM, Macdonald RR, Major KA, *et al*.

No. 43
The effectiveness and cost-effectiveness of dual-chamber pacemakers compared with single-chamber pacemakers for bradycardia due to atrioventricular block or sick sinus syndrome: systematic review and economic evaluation.
By Castelnuovo E, Stein K, Pitt M, Garside R, Payne E.

No. 44
Newborn screening for congenital heart defects: a systematic review and cost-effectiveness analysis.
By Knowles R, Griebsch I, Dezateux C, Brown J, Bull C, Wren C.

No. 45
The clinical and cost-effectiveness of left ventricular assist devices for end-stage heart failure: a systematic review and economic evaluation.
By Clegg AJ, Scott DA, Loveman E, Colquitt J, Hutchinson J, Royle P, *et al*.

No. 46
The effectiveness of the Heidelberg Retina Tomograph and laser diagnostic glaucoma scanning system (GDx) in detecting and monitoring glaucoma.
By Kwartz AJ, Henson DB, Harper RA, Spencer AF, McLeod D.

No. 47
Clinical and cost-effectiveness of autologous chondrocyte implantation for cartilage defects in knee joints: systematic review and economic evaluation.
By Clar C, Cummins E, McIntyre L, Thomas S, Lamb J, Bain L, *et al*.

No. 48
Systematic review of effectiveness of different treatments for childhood retinoblastoma.
By McDaid C, Hartley S, Bagnall A-M, Ritchie G, Light K, Riemsma R.

No. 49
Towards evidence-based guidelines for the prevention of venous thromboembolism: systematic reviews of mechanical methods, oral anticoagulation, dextran and regional anaesthesia as thromboprophylaxis.
By Roderick P, Ferris G, Wilson K, Halls H, Jackson D, Collins R, *et al*.

No. 50
The effectiveness and cost-effectiveness of parent training/education programmes for the treatment of conduct disorder, including oppositional defiant disorder, in children.
By Dretzke J, Frew E, Davenport C, Barlow J, Stewart-Brown S, Sandercock J, *et al*.

Volume 10, 2006

No. 1
The clinical and cost-effectiveness of donepezil, rivastigmine, galantamine and memantine for Alzheimer's disease.
By Loveman E, Green C, Kirby J, Takeda A, Picot J, Payne E, *et al*.

No. 2
FOOD: a multicentre randomised trial evaluating feeding policies in patients admitted to hospital with a recent stroke.
By Dennis M, Lewis S, Cranswick G, Forbes J.

No. 3
The clinical effectiveness and cost-effectiveness of computed tomography screening for lung cancer: systematic reviews.
By Black C, Bagust A, Boland A, Walker S, McLeod C, De Verteuil R, *et al*.

No. 4
A systematic review of the effectiveness and cost-effectiveness of neuroimaging assessments used to visualise the seizure focus in people with refractory epilepsy being considered for surgery.
By Whiting P, Gupta R, Burch J, Mujica Mota RE, Wright K, Marson A, *et al*.

No. 5
Comparison of conference abstracts and presentations with full-text articles in the health technology assessments of rapidly evolving technologies.
By Dundar Y, Dodd S, Dickson R, Walley T, Haycox A, Williamson PR.

No. 6
Systematic review and evaluation of methods of assessing urinary incontinence.
By Martin JL, Williams KS, Abrams KR, Turner DA, Sutton AJ, Chapple C, *et al*.

No. 7
The clinical effectiveness and cost-effectiveness of newer drugs for children with epilepsy. A systematic review.
By Connock M, Frew E, Evans B-W, Bryan S, Cummins C, Fry-Smith A, *et al*.

No. 8
Surveillance of Barrett's oesophagus: exploring the uncertainty through systematic review, expert workshop and economic modelling.
By Garside R, Pitt M, Somerville M, Stein K, Price A, Gilbert N.

No. 9
Topotecan, pegylated liposomal doxorubicin hydrochloride and paclitaxel for second-line or subsequent treatment of advanced ovarian cancer: a systematic review and economic evaluation.
By Main C, Bojke L, Griffin S, Norman G, Barbieri M, Mather L, *et al*.

No. 10
Evaluation of molecular techniques in prediction and diagnosis of cytomegalovirus disease in immunocompromised patients.
By Szczepura A, Westmoreland D, Vinogradova Y, Fox J, Clark M.

No. 11
Screening for thrombophilia in high-risk situations: systematic review and cost-effectiveness analysis. The Thrombosis: Risk and Economic Assessment of Thrombophilia Screening (TREATS) study.
By Wu O, Robertson L, Twaddle S, Lowe GDO, Clark P, Greaves M, *et al*.

No. 12
A series of systematic reviews to inform a decision analysis for sampling and treating infected diabetic foot ulcers.
By Nelson EA, O'Meara S, Craig D, Iglesias C, Golder S, Dalton J, *et al*.

No. 13
Randomised clinical trial, observational study and assessment of cost-effectiveness of the treatment of varicose veins (REACTIV trial).
By Michaels JA, Campbell WB, Brazier JE, MacIntyre JB, Palfreyman SJ, Ratcliffe J, *et al*.

No. 14
The cost-effectiveness of screening for oral cancer in primary care.
By Speight PM, Palmer S, Moles DR, Downer MC, Smith DH, Henriksson M *et al*.

No. 15
Measurement of the clinical and cost-effectiveness of non-invasive diagnostic testing strategies for deep vein thrombosis.
By Goodacre S, Sampson F, Stevenson M, Wailoo A, Sutton A, Thomas S, *et al*.

No. 16
Systematic review of the effectiveness and cost-effectiveness of HealOzone® for the treatment of occlusal pit/fissure caries and root caries.
By Brazzelli M, McKenzie L, Fielding S, Fraser C, Clarkson J, Kilonzo M, *et al*.

No. 17
Randomised controlled trials of conventional antipsychotic versus new atypical drugs, and new atypical drugs versus clozapine, in people with schizophrenia responding poorly to, or intolerant of, current drug treatment.
By Lewis SW, Davies L, Jones PB, Barnes TRE, Murray RM, Kerwin R, *et al*.

No. 18
Diagnostic tests and algorithms used in the investigation of haematuria: systematic reviews and economic evaluation.
By Rodgers M, Nixon J, Hempel S, Aho T, Kelly J, Neal D, *et al*.

No. 19
Cognitive behavioural therapy in addition to antispasmodic therapy for irritable bowel syndrome in primary care: randomised controlled trial.
By Kennedy TM, Chalder T, McCrone P, Darnley S, Knapp M, Jones RH, *et al*.

No. 20
A systematic review of the clinical effectiveness and cost-effectiveness of enzyme replacement therapies for Fabry's disease and mucopolysaccharidosis type 1.
By Connock M, Juarez-Garcia A, Frew E, Mans A, Dretzke J, Fry-Smith A, *et al*.

No. 21
Health benefits of antiviral therapy for mild chronic hepatitis C: randomised controlled trial and economic evaluation.
By Wright M, Grieve R, Roberts J, Main J, Thomas HC on behalf of the UK Mild Hepatitis C Trial Investigators.

No. 22
Pressure relieving support surfaces: a randomised evaluation.
By Nixon J, Nelson EA, Cranny G, Iglesias CP, Hawkins K, Cullum NA, *et al*.

No. 23
A systematic review and economic model of the effectiveness and cost-effectiveness of methylphenidate, dexamfetamine and atomoxetine for the treatment of attention deficit hyperactivity disorder in children and adolescents.
By King S, Griffin S, Hodges Z, Weatherly H, Asseburg C, Richardson G, *et al*.

No. 24
The clinical effectiveness and cost-effectiveness of enzyme replacement therapy for Gaucher's disease: a systematic review.
By Connock M, Burls A, Frew E, Fry-Smith A, Juarez-Garcia A, McCabe C, *et al*.

No. 25
Effectiveness and cost-effectiveness of salicylic acid and cryotherapy for cutaneous warts. An economic decision model.
By Thomas KS, Keogh-Brown MR, Chalmers JR, Fordham RJ, Holland RC, Armstrong SJ, *et al*.

No. 26
A systematic literature review of the effectiveness of non-pharmacological interventions to prevent wandering in dementia and evaluation of the ethical implications and acceptability of their use.
By Robinson L, Hutchings D, Corner L, Beyer F, Dickinson H, Vanoli A, *et al*.

No. 27
A review of the evidence on the effects and costs of implantable cardioverter defibrillator therapy in different patient groups, and modelling of cost-effectiveness and cost–utility for these groups in a UK context.
By Buxton M, Caine N, Chase D, Connelly D, Grace A, Jackson C, *et al*.

No. 28
Adefovir dipivoxil and pegylated interferon alfa-2a for the treatment of chronic hepatitis B: a systematic review and economic evaluation.
By Shepherd J, Jones J, Takeda A, Davidson P, Price A.

No. 29
An evaluation of the clinical and cost-effectiveness of pulmonary artery catheters in patient management in intensive care: a systematic review and a randomised controlled trial.
By Harvey S, Stevens K, Harrison D, Young D, Brampton W, McCabe C, *et al*.

No. 30
Accurate, practical and cost-effective assessment of carotid stenosis in the UK.
By Wardlaw JM, Chappell FM, Stevenson M, De Nigris E, Thomas S, Gillard J, *et al*.

No. 31
Etanercept and infliximab for the treatment of psoriatic arthritis: a systematic review and economic evaluation.
By Woolacott N, Bravo Vergel Y, Hawkins N, Kainth A, Khadjesari Z, Misso K, *et al*.

No. 32
The cost-effectiveness of testing for hepatitis C in former injecting drug users.
By Castelnuovo E, Thompson-Coon J, Pitt M, Cramp M, Siebert U, Price A, *et al*.

No. 33
Computerised cognitive behaviour therapy for depression and anxiety update: a systematic review and economic evaluation.
By Kaltenthaler E, Brazier J, De Nigris E, Tumur I, Ferriter M, Beverley C, *et al*.

No. 34
Cost-effectiveness of using prognostic information to select women with breast cancer for adjuvant systemic therapy.
By Williams C, Brunskill S, Altman D, Briggs A, Campbell H, Clarke M, *et al*.

No. 35
Psychological therapies including dialectical behaviour therapy for borderline personality disorder: a systematic review and preliminary economic evaluation.
By Brazier J, Tumur I, Holmes M, Ferriter M, Parry G, Dent-Brown K, *et al*.

No. 36
Clinical effectiveness and cost-effectiveness of tests for the diagnosis and investigation of urinary tract infection in children: a systematic review and economic model.
By Whiting P, Westwood M, Bojke L, Palmer S, Richardson G, Cooper J, *et al*.

No. 37
Cognitive behavioural therapy in chronic fatigue syndrome: a randomised controlled trial of an outpatient group programme.
By O'Dowd H, Gladwell P, Rogers CA, Hollinghurst S, Gregory A.

No. 38
A comparison of the cost-effectiveness of five strategies for the prevention of non-steroidal anti-inflammatory drug-induced gastrointestinal toxicity: a systematic review with economic modelling.
By Brown TJ, Hooper L, Elliott RA, Payne K, Webb R, Roberts C, *et al*.

No. 39
The effectiveness and cost-effectiveness of computed tomography screening for coronary artery disease: systematic review.
By Waugh N, Black C, Walker S, McIntyre L, Cummins E, Hillis G.

No. 40
What are the clinical outcome and cost-effectiveness of endoscopy undertaken by nurses when compared with doctors? A Multi-Institution Nurse Endoscopy Trial (MINuET).
By Williams J, Russell I, Durai D, Cheung W-Y, Farrin A, Bloor K, *et al*.

No. 41
The clinical and cost-effectiveness of oxaliplatin and capecitabine for the adjuvant treatment of colon cancer: systematic review and economic evaluation.
By Pandor A, Eggington S, Paisley S, Tappenden P, Sutcliffe P.

No. 42
A systematic review of the effectiveness of adalimumab, etanercept and infliximab for the treatment of rheumatoid arthritis in adults and an economic evaluation of their cost-effectiveness.
By Chen Y-F, Jobanputra P, Barton P, Jowett S, Bryan S, Clark W, *et al*.

No. 43
Telemedicine in dermatology: a randomised controlled trial.
By Bowns IR, Collins K, Walters SJ, McDonagh AJG.

No. 44
Cost-effectiveness of cell salvage and alternative methods of minimising perioperative allogeneic blood transfusion: a systematic review and economic model.
By Davies L, Brown TJ, Haynes S, Payne K, Elliott RA, McCollum C.

No. 45
Clinical effectiveness and cost-effectiveness of laparoscopic surgery for colorectal cancer: systematic reviews and economic evaluation.
By Murray A, Lourenco T, de Verteuil R, Hernandez R, Fraser C, McKinley A, *et al*.

No. 46
Etanercept and efalizumab for the treatment of psoriasis: a systematic review.
By Woolacott N, Hawkins N, Mason A, Kainth A, Khadjesari Z, Bravo Vergel Y, *et al*.

No. 47
Systematic reviews of clinical decision tools for acute abdominal pain.
By Liu JLY, Wyatt JC, Deeks JJ, Clamp S, Keen J, Verde P, *et al*.

No. 48
Evaluation of the ventricular assist device programme in the UK.
By Sharples L, Buxton M, Caine N, Cafferty F, Demiris N, Dyer M, *et al*.

No. 49
A systematic review and economic model of the clinical and cost-effectiveness of immunosuppressive therapy for renal transplantation in children.
By Yao G, Albon E, Adi Y, Milford D, Bayliss S, Ready A, *et al*.

No. 50
Amniocentesis results: investigation of anxiety. The ARIA trial.
By Hewison J, Nixon J, Fountain J, Cocks K, Jones C, Mason G, *et al*.

Volume 11, 2007

No. 1
Pemetrexed disodium for the treatment of malignant pleural mesothelioma: a systematic review and economic evaluation.
By Dundar Y, Bagust A, Dickson R, Dodd S, Green J, Haycox A, *et al*.

No. 2
A systematic review and economic model of the clinical effectiveness and cost-effectiveness of docetaxel in combination with prednisone or prednisolone for the treatment of hormone-refractory metastatic prostate cancer.
By Collins R, Fenwick E, Trowman R, Perard R, Norman G, Light K, *et al*.

No. 3
A systematic review of rapid diagnostic tests for the detection of tuberculosis infection.
By Dinnes J, Deeks J, Kunst H, Gibson A, Cummins E, Waugh N, *et al*.

No. 4
The clinical effectiveness and cost-effectiveness of strontium ranelate for the prevention of osteoporotic fragility fractures in postmenopausal women.
By Stevenson M, Davis S, Lloyd-Jones M, Beverley C.

No. 5
A systematic review of quantitative and qualitative research on the role and effectiveness of written information available to patients about individual medicines.
By Raynor DK, Blenkinsopp A, Knapp P, Grime J, Nicolson DJ, Pollock K, *et al*.

No. 6
Oral naltrexone as a treatment for relapse prevention in formerly opioid-dependent drug users: a systematic review and economic evaluation.
By Adi Y, Juarez-Garcia A, Wang D, Jowett S, Frew E, Day E, *et al*.

No. 7
Glucocorticoid-induced osteoporosis: a systematic review and cost–utility analysis.
By Kanis JA, Stevenson M, McCloskey EV, Davis S, Lloyd-Jones M.

No. 8
Epidemiological, social, diagnostic and economic evaluation of population screening for genital chlamydial infection.
By Low N, McCarthy A, Macleod J, Salisbury C, Campbell R, Roberts TE, *et al*.

No. 9
Methadone and buprenorphine for the management of opioid dependence: a systematic review and economic evaluation.
By Connock M, Juarez-Garcia A, Jowett S, Frew E, Liu Z, Taylor RJ, *et al*.

No. 10
Exercise Evaluation Randomised Trial (EXERT): a randomised trial comparing GP referral for leisure centre-based exercise, community-based walking and advice only.
By Isaacs AJ, Critchley JA, See Tai S, Buckingham K, Westley D, Harridge SDR, *et al*.

No. 11
Interferon alfa (pegylated and non-pegylated) and ribavirin for the treatment of mild chronic hepatitis C: a systematic review and economic evaluation.
By Shepherd J, Jones J, Hartwell D, Davidson P, Price A, Waugh N.

No. 12
Systematic review and economic evaluation of bevacizumab and cetuximab for the treatment of metastatic colorectal cancer.
By Tappenden P, Jones R, Paisley S, Carroll C.

No. 13
A systematic review and economic evaluation of epoetin alfa, epoetin beta and darbepoetin alfa in anaemia associated with cancer, especially that attributable to cancer treatment.
By Wilson J, Yao GL, Raftery J, Bohlius J, Brunskill S, Sandercock J, *et al*.

No. 14
A systematic review and economic evaluation of statins for the prevention of coronary events.
By Ward S, Lloyd Jones M, Pandor A, Holmes M, Ara R, Ryan A, *et al*.

No. 15
A systematic review of the effectiveness and cost-effectiveness of different models of community-based respite care for frail older people and their carers.
By Mason A, Weatherly H, Spilsbury K, Arksey H, Golder S, Adamson J, *et al*.

No. 16
Additional therapy for young children with spastic cerebral palsy: a randomised controlled trial.
By Weindling AM, Cunningham CC, Glenn SM, Edwards RT, Reeves DJ.

No. 17
Screening for type 2 diabetes: literature review and economic modelling.
By Waugh N, Scotland G, McNamee P, Gillett M, Brennan A, Goyder E, *et al*.

No. 18
The effectiveness and cost-effectiveness of cinacalcet for secondary hyperparathyroidism in end-stage renal disease patients on dialysis: a systematic review and economic evaluation.
By Garside R, Pitt M, Anderson R, Mealing S, Roome C, Snaith A, *et al*.

No. 19
The clinical effectiveness and cost-effectiveness of gemcitabine for metastatic breast cancer: a systematic review and economic evaluation.
By Takeda AL, Jones J, Loveman E, Tan SC, Clegg AJ.

No. 20
A systematic review of duplex ultrasound, magnetic resonance angiography and computed tomography angiography for the diagnosis and assessment of symptomatic, lower limb peripheral arterial disease.
By Collins R, Cranny G, Burch J, Aguiar-Ibáñez R, Craig D, Wright K, *et al*.

No. 21
The clinical effectiveness and cost-effectiveness of treatments for children with idiopathic steroid-resistant nephrotic syndrome: a systematic review.
By Colquitt JL, Kirby J, Green C, Cooper K, Trompeter RS.

No. 22
A systematic review of the routine monitoring of growth in children of primary school age to identify growth-related conditions.
By Fayter D, Nixon J, Hartley S, Rithalia A, Butler G, Rudolf M, *et al*.

No. 23
Systematic review of the effectiveness of preventing and treating *Staphylococcus aureus* carriage in reducing peritoneal catheter-related infections.
By McCormack K, Rabindranath K, Kilonzo M, Vale L, Fraser C, McIntyre L, *et al*.

No. 24
The clinical effectiveness and cost of repetitive transcranial magnetic stimulation versus electroconvulsive therapy in severe depression: a multicentre pragmatic randomised controlled trial and economic analysis.
By McLoughlin DM, Mogg A, Eranti S, Pluck G, Purvis R, Edwards D, *et al*.

No. 25
A randomised controlled trial and economic evaluation of direct versus indirect and individual versus group modes of speech and language therapy for children with primary language impairment.
By Boyle J, McCartney E, Forbes J, O'Hare A.

No. 26
Hormonal therapies for early breast cancer: systematic review and economic evaluation.
By Hind D, Ward S, De Nigris E, Simpson E, Carroll C, Wyld L.

No. 27
Cardioprotection against the toxic effects of anthracyclines given to children with cancer: a systematic review.
By Bryant J, Picot J, Levitt G, Sullivan I, Baxter L, Clegg A.

No. 28
Adalimumab, etanercept and infliximab for the treatment of ankylosing spondylitis: a systematic review and economic evaluation.
By McLeod C, Bagust A, Boland A, Dagenais P, Dickson R, Dundar Y, *et al*.

No. 29
Prenatal screening and treatment strategies to prevent group B streptococcal and other bacterial infections in early infancy: cost-effectiveness and expected value of information analyses.
By Colbourn T, Asseburg C, Bojke L, Philips Z, Claxton K, Ades AE, *et al*.

No. 30
Clinical effectiveness and cost-effectiveness of bone morphogenetic proteins in the non-healing of fractures and spinal fusion: a systematic review.
By Garrison KR, Donell S, Ryder J, Shemilt I, Mugford M, Harvey I, *et al*.

No. 31
A randomised controlled trial of postoperative radiotherapy following breast-conserving surgery in a minimum-risk older population. The PRIME trial.
By Prescott RJ, Kunkler IH, Williams LJ, King CC, Jack W, van der Pol M, *et al*.

No. 32
Current practice, accuracy, effectiveness and cost-effectiveness of the school entry hearing screen.
By Bamford J, Fortnum H, Bristow K, Smith J, Vamvakas G, Davies L, *et al*.

No. 33
The clinical effectiveness and cost-effectiveness of inhaled insulin in diabetes mellitus: a systematic review and economic evaluation.
By Black C, Cummins E, Royle P, Philip S, Waugh N.

No. 34
Surveillance of cirrhosis for hepatocellular carcinoma: systematic review and economic analysis.
By Thompson Coon J, Rogers G, Hewson P, Wright D, Anderson R, Cramp M, *et al*.

No. 35
The Birmingham Rehabilitation Uptake Maximisation Study (BRUM). Home-based compared with hospital-based cardiac rehabilitation in a multi-ethnic population: cost-effectiveness and patient adherence.
By Jolly K, Taylor R, Lip GYH, Greenfield S, Raftery J, Mant J, *et al*.

No. 36
A systematic review of the clinical, public health and cost-effectiveness of rapid diagnostic tests for the detection and identification of bacterial intestinal pathogens in faeces and food.
By Abubakar I, Irvine L, Aldus CF, Wyatt GM, Fordham R, Schelenz S, *et al*.

No. 37
A randomised controlled trial examining the longer-term outcomes of standard versus new antiepileptic drugs. The SANAD trial.
By Marson AG, Appleton R, Baker GA, Chadwick DW, Doughty J, Eaton B, *et al*.

No. 38
Clinical effectiveness and cost-effectiveness of different models of managing long-term oral anticoagulation therapy: a systematic review and economic modelling.
By Connock M, Stevens C, Fry-Smith A, Jowett S, Fitzmaurice D, Moore D, *et al*.

No. 39
A systematic review and economic model of the clinical effectiveness and cost-effectiveness of interventions for preventing relapse in people with bipolar disorder.
By Soares-Weiser K, Bravo Vergel Y, Beynon S, Dunn G, Barbieri M, Duffy S, *et al*.

No. 40
Taxanes for the adjuvant treatment of early breast cancer: systematic review and economic evaluation.
By Ward S, Simpson E, Davis S, Hind D, Rees A, Wilkinson A.

No. 41
The clinical effectiveness and cost-effectiveness of screening for open angle glaucoma: a systematic review and economic evaluation.
By Burr JM, Mowatt G, Hernández R, Siddiqui MAR, Cook J, Lourenco T, *et al*.

No. 42
Acceptability, benefit and costs of early screening for hearing disability: a study of potential screening tests and models.
By Davis A, Smith P, Ferguson M, Stephens D, Gianopoulos I.

No. 43
Contamination in trials of educational interventions.
By Keogh-Brown MR, Bachmann MO, Shepstone L, Hewitt C, Howe A, Ramsay CR, *et al*.

No. 44
Overview of the clinical effectiveness of positron emission tomography imaging in selected cancers.
By Facey K, Bradbury I, Laking G, Payne E.

No. 45
The effectiveness and cost-effectiveness of carmustine implants and temozolomide for the treatment of newly diagnosed high-grade glioma: a systematic review and economic evaluation.
By Garside R, Pitt M, Anderson R, Rogers G, Dyer M, Mealing S, *et al*.

No. 46
Drug-eluting stents: a systematic review and economic evaluation.
By Hill RA, Boland A, Dickson R, Dündar Y, Haycox A, McLeod C, *et al*.

No. 47
The clinical effectiveness and cost-effectiveness of cardiac resynchronisation (biventricular pacing) for heart failure: systematic review and economic model.
By Fox M, Mealing S, Anderson R, Dean J, Stein K, Price A, *et al*.

No. 48
Recruitment to randomised trials: strategies for trial enrolment and participation study. The STEPS study.
By Campbell MK, Snowdon C, Francis D, Elbourne D, McDonald AM, Knight R, *et al*.

No. 49
Cost-effectiveness of functional cardiac testing in the diagnosis and management of coronary artery disease: a randomised controlled trial. The CECaT trial.
By Sharples L, Hughes V, Crean A, Dyer M, Buxton M, Goldsmith K, *et al*.

No. 50
Evaluation of diagnostic tests when there is no gold standard. A review of methods.
By Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PMM.

Health Technology Assessment Programme

Director,
Professor Tom Walley,
Director, NHS HTA Programme,
Department of Pharmacology & Therapeutics,
University of Liverpool

Deputy Director,
Professor Jon Nicholl,
Director, Medical Care Research Unit, University of Sheffield,
School of Health and Related Research

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 Robin E Ferner,
Consultant Physician and Director, West Midlands Centre for Adverse Drug Reactions,
City Hospital NHS Trust,
Birmingham

Dr Edmund Jessop, Medical Adviser, National Specialist, Commissioning Advisory Group (NSCAG), Department of Health, London

Professor Jon Nicholl, Director, Medical Care Research Unit,
University of Sheffield,
School of Health and Related Research

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 Jon Nicholl,
Director, Medical Care Research Unit, University of Sheffield,
School of Health and Related Research

Deputy Chair,
Dr Andrew Farmer,
University Lecturer in General Practice, Department of Primary Health Care,
University of Oxford

Dr Jeffrey Aronson,
Reader in Clinical Pharmacology, Department of Clinical Pharmacology,
Radcliffe Infirmary, Oxford

Professor Deborah Ashby,
Professor of Medical Statistics,
Department of Environmental and Preventative Medicine,
Queen Mary University of London

Professor Ann Bowling,
Professor of Health Services Research, Primary Care and Population Studies,
University College London

Professor John Cairns,
Professor of Health Economics,
Public Health Policy,
London School of Hygiene and Tropical Medicine,
London

Professor Nicky Cullum,
Director of Centre for Evidence Based Nursing, Department of Health Sciences, University of York

Professor Jon Deeks,
Professor of Health Statistics,
University of Birmingham

Professor Jenny Donovan,
Professor of Social Medicine,
Department of Social Medicine,
University of Bristol

Professor Freddie Hamdy,
Professor of Urology,
University of Sheffield

Professor Allan House,
Professor of Liaison Psychiatry,
University of Leeds

Professor Sallie Lamb, Director,
Warwick Clinical Trials Unit,
University of Warwick

Professor Stuart Logan,
Director of Health & Social Care Research, The Peninsula Medical School, Universities of Exeter & Plymouth

Professor Miranda Mugford,
Professor of Health Economics,
University of East Anglia

Dr Linda Patterson,
Consultant Physician,
Department of Medicine,
Burnley General Hospital

Professor Ian Roberts,
Professor of Epidemiology & Public Health, Intervention Research Unit, London School of Hygiene and Tropical Medicine

Professor Mark Sculpher,
Professor of Health Economics,
Centre for Health Economics,
Institute for Research in the Social Services,
University of York

Professor Kate Thomas,
Professor of Complementary and Alternative Medicine,
University of Leeds

Professor David John Torgerson,
Director of York Trial Unit,
Department of Health Sciences,
University of York

Professor Hywel Williams,
Professor of Dermato-Epidemiology,
University of Nottingham

Current and past membership details of all HTA 'committees' are available from the HTA website (www.hta.ac.uk)

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 of 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

Ms Dea Birkett, Service User Representative, London

Dr Paul Cockcroft, Consultant Medical Microbiologist and Clinical Director of Pathology, Department of Clinical Microbiology, St Mary's Hospital, Portsmouth

Professor Adrian K Dixon, Professor of Radiology, University Department of Radiology, University of Cambridge Clinical School

Dr David Elliman, Consultant in Community Child Health, Islington PCT & Great Ormond Street Hospital, London

Professor Glyn Elwyn, Research Chair, Centre for Health Sciences Research, Cardiff University, Department of General Practice, Cardiff

Professor Paul Glasziou, Director, Centre for Evidence-Based Practice, University of Oxford

Dr Jennifer J Kurinczuk, Consultant Clinical Epidemiologist, National Perinatal Epidemiology Unit, Oxford

Dr Susanne M Ludgate, Clinical Director, Medicines & Healthcare Products Regulatory Agency, London

Mr Stephen Pilling, Director, Centre for Outcomes, Research & Effectiveness, Joint Director, National Collaborating Centre for Mental Health, University College London

Mrs Una Rennard, Service User Representative, Oxford

Dr Phil Shackley, Senior Lecturer in Health Economics, Academic Vascular Unit, University of Sheffield

Dr Margaret Somerville, Director of Public Health Learning, Peninsula Medical School, University of Plymouth

Dr Graham Taylor, Scientific Director & Senior Lecturer, Regional DNA Laboratory, The Leeds Teaching Hospitals

Professor Lindsay Wilson Turnbull, Scientific Director, Centre for MR Investigations & YCR Professor of Radiology, University of Hull

Professor Martin J Whittle, Clinical Co-director, National Co-ordinating Centre for Women's and Childhealth

Dr Dennis Wright, Consultant Biochemist & Clinical Director, The North West London Hospitals NHS Trust, Middlesex

Pharmaceuticals Panel

Members

Chair,
Professor Robin Ferner, Consultant Physician and Director, West Midlands Centre for Adverse Drug Reactions, City Hospital NHS Trust, Birmingham

Ms Anne Baileff, Consultant Nurse in First Contact Care, Southampton City Primary Care Trust, University of Southampton

Professor Imti Choonara, Professor in Child Health, Academic Division of Child Health, University of Nottingham

Professor John Geddes, Professor of Epidemiological Psychiatry, University of Oxford

Mrs Barbara Greggains, Non-Executive Director, Greggains Management Ltd

Dr Bill Gutteridge, Medical Adviser, National Specialist Commissioning Advisory Group (NSCAG), London

Mrs Sharon Hart, Consultant Pharmaceutical Adviser, Reading

Dr Jonathan Karnon, Senior Research Fellow, Health Economics and Decision Science, University of Sheffield

Dr Yoon Loke, Senior Lecturer in Clinical Pharmacology, University of East Anglia

Ms Barbara Meredith, Lay Member, Epsom

Dr Andrew Prentice, Senior Lecturer and Consultant Obstetrician & Gynaecologist, Department of Obstetrics & Gynaecology, University of Cambridge

Dr Frances Rotblat, CPMP Delegate, Medicines & Healthcare Products Regulatory Agency, London

Dr Martin Shelly, General Practitioner, Leeds

Mrs Katrina Simister, Assistant Director New Medicines, National Prescribing Centre, Liverpool

Dr Richard Tiner, Medical Director, Medical Department, Association of the British Pharmaceutical Industry, London

Therapeutic Procedures Panel

Members

Chair,
Professor Bruce Campbell, Consultant Vascular and General Surgeon, Department of Surgery, Royal Devon & Exeter Hospital

Dr Mahmood Adil, Deputy Regional Director of Public Health, Department of Health, Manchester

Dr Aileen Clarke, Consultant in Public Health, Public Health Resource Unit, Oxford

Professor Matthew Cooke, Professor of Emergency Medicine, Warwick Emergency Care and Rehabilitation, University of Warwick

Mr Mark Emberton, Senior Lecturer in Oncological Urology, Institute of Urology, University College Hospital

Professor Paul Gregg, Professor of Orthopaedic Surgical Science, Department of General Practice and Primary Care, South Tees Hospital NHS Trust, Middlesbrough

Ms Maryann L Hardy, Lecturer, Division of Radiography, University of Bradford

Dr Simon de Lusignan, Senior Lecturer, Primary Care Informatics, Department of Community Health Sciences, St George's Hospital Medical School, London

Dr Peter Martin, Consultant Neurologist, Addenbrooke's Hospital, Cambridge

Professor Neil McIntosh, Edward Clark Professor of Child Life & Health, Department of Child Life & Health, University of Edinburgh

Professor Jim Neilson, Professor of Obstetrics and Gynaecology, Department of Obstetrics and Gynaecology, University of Liverpool

Dr John C Pounsford, Consultant Physician, Directorate of Medical Services, North Bristol NHS Trust

Dr Karen Roberts, Nurse Consultant, Queen Elizabeth Hospital, Gateshead

Dr Vimal Sharma, Consultant Psychiatrist/Hon. Senior Lecturer, Mental Health Resource Centre, Cheshire and Wirral Partnership NHS Trust, Wallasey

Professor Scott Weich, Professor of Psychiatry, Division of Health in the Community, University of Warwick

Disease Prevention Panel

Members

Chair,
Dr Edmund Jessop, Medical Adviser, National Specialist Commissioning Advisory Group (NSCAG), London

Mrs Sheila Clark, Chief Executive, St James's Hospital, Portsmouth

Mr Richard Copeland, Lead Pharmacist: Clinical Economy/Interface, Wansbeck General Hospital, Northumberland

Dr Elizabeth Fellow-Smith, Medical Director, West London Mental Health Trust, Middlesex

Mr Ian Flack, Director PPI Forum Support, Council of Ethnic Minority Voluntary Sector Organisations, Stratford

Dr John Jackson, General Practitioner, Newcastle upon Tyne

Mrs Veronica James, Chief Officer, Horsham District Age Concern, Horsham

Professor Mike Kelly, Director, Centre for Public Health Excellence, National Institute for Health and Clinical Excellence, London

Professor Yi Mien Koh, Director of Public Health and Medical Director, London NHS (North West London Strategic Health Authority), London

Ms Jeanett Martin, Director of Clinical Leadership & Quality, Lewisham PCT, London

Dr Chris McCall, General Practitioner, Dorset

Dr David Pencheon, Director, Eastern Region Public Health Observatory, Cambridge

Dr Ken Stein, Senior Clinical Lecturer in Public Health, Director, Peninsula Technology Assessment Group, University of Exeter, Exeter

Dr Carol Tannahill, Director, Glasgow Centre for Population Health, Glasgow

Professor Margaret Thorogood, Professor of Epidemiology, University of Warwick, Coventry

Dr Ewan Wilkinson, Consultant in Public Health, Royal Liverpool University Hospital, Liverpool

Expert Advisory Network

Members

Professor Douglas Altman, Professor of Statistics in Medicine, Centre for Statistics in Medicine, University of Oxford

Professor John Bond, Director, Centre for Health Services Research, University of Newcastle upon Tyne, School of Population & Health Sciences, Newcastle upon Tyne

Professor Andrew Bradbury, Professor of Vascular Surgery, Solihull Hospital, Birmingham

Mr Shaun Brogan, Chief Executive, Ridgeway Primary Care Group, Aylesbury

Mrs Stella Burnside OBE, Chief Executive, Regulation and Improvement Authority, Belfast

Ms Tracy Bury, Project Manager, World Confederation for Physical Therapy, London

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 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

Dr Carl Counsell, Clinical Senior Lecturer in Neurology, Department of Medicine & Therapeutics, University of Aberdeen

Professor Howard Cuckle, Professor of Reproductive Epidemiology, Department of Paediatrics, Obstetrics & Gynaecology, University of Leeds

Dr Katherine Darton, Information Unit, MIND – The Mental Health Charity, London

Professor Carol Dezateux, Professor of Paediatric Epidemiology, London

Dr Keith Dodd, Consultant Paediatrician, Derby

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

Professor Gene Feder, Professor of Primary Care Research & Development, Centre for Health Sciences, Barts & The London Queen Mary's School of Medicine & Dentistry, London

Mr Leonard R Fenwick, Chief Executive, Newcastle upon Tyne Hospitals 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

Dr Neville Goodman, Consultant Anaesthetist, Southmead Hospital, Bristol

Professor Robert E Hawkins, CRC Professor and Director of Medical Oncology, Christie CRC Research Centre, Christie Hospital NHS Trust, Manchester

Professor Allen Hutchinson, Director of Public Health & Deputy Dean of ScHARR, Department of Public Health, University of Sheffield

Professor Peter Jones, Professor of Psychiatry, University of Cambridge, Cambridge

Professor Stan Kaye, Cancer Research UK Professor of Medical Oncology, Section of Medicine, Royal Marsden Hospital & Institute of Cancer Research, Surrey

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 Julian Little, Professor of Human Genome Epidemiology, Department of Epidemiology & Community Medicine, University of Ottawa

Professor Rajan Madhok, Consultant in Public Health, South Manchester Primary Care Trust, Manchester

Professor Alexander Markham, Director, Molecular Medicine Unit, St James's University Hospital, Leeds

Professor Alistaire McGuire, Professor of Health Economics, London School of Economics

Dr Peter Moore, Freelance Science Writer, Ashtead

Dr Andrew Mortimore, Public Health Director, Southampton City Primary Care Trust, Southampton

Dr Sue Moss, Associate Director, Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton

Mrs Julietta Patnick, Director, NHS Cancer Screening Programmes, Sheffield

Professor Robert Peveler, Professor of Liaison Psychiatry, Royal South Hants Hospital, Southampton

Professor Chris Price, Visiting Professor in Clinical Biochemistry, University of Oxford

Professor William Rosenberg, Professor of Hepatology and Consultant Physician, University of Southampton, Southampton

Professor Peter Sandercock, Professor of Medical Neurology, Department of Clinical Neurosciences, University of Edinburgh

Dr Susan Schonfield, Consultant in Public Health, Hillingdon PCT, Middlesex

Dr Eamonn Sheridan, Consultant in Clinical Genetics, Genetics Department, St James's University Hospital, Leeds

Professor Sarah Stewart-Brown, Professor of Public Health, University of Warwick, Division of Health in the Community Warwick Medical School, LWMS, Coventry

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.hta.ac.uk) 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@hta.ac.uk
http://www.hta.ac.uk

ISSN 1366-5278