



UvA-DARE (Digital Academic Repository)

Systematic reviews of diagnostic test accuracy

Leeflang, M.M.G.

Publication date

2008

Document Version

Final published version

[Link to publication](#)

Citation for published version (APA):

Leeflang, M. M. G. (2008). *Systematic reviews of diagnostic test accuracy*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Systematic Reviews of Diagnostic Test Accuracy



Mariska Leeflang

Systematic reviews of diagnostic test accuracy

Systematic reviews of diagnostic test accuracy
Thesis, University of Amsterdam, The Netherlands

This thesis was prepared at the Department of Clinical Epidemiology, Biostatistics and Bioinformatics at the Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands.

The publication was financially supported by The Dutch Cochrane Centre, Stichting tot Bevordering van de Klinische Epidemiologie and the Academic Medical Center.

ISBN: 978-90-9023139-6

No part of this thesis may be reproduced, stored or transmitted in any way and by no means, without permission of the author. A digital version of this thesis can be found at <http://dare.uva.nl>.

Cover design: Mariska Leeftang
Layout: Fay Koen Tjoa
Printed by: Grafisch Bedrijf Ponsen & Looijen b.v.

About the cover: the cover is inspired by Celtic artwork and reflects the process of a systematic review: parts become a whole. The anthropomorphic (human-like) and zoomorphic (animal-like) creatures represent the background of the author. The stethoscopes and the corners refer specifically to diagnostic test accuracy reviews. The snakes eating their own tail stand in Celtic mythology for longevity and the ever-lasting life cycle.

Systematic reviews of diagnostic test accuracy

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. D.C. van den Boom,

ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 1 juli 2008, te 14:00 uur

door

Maria Mariska Geertruida Leeflang

geboren te Edam-Volendam

Promotiecommissie:

Promotor: Prof. dr. P.M.M. Bossuyt

Co-promotores: Dr. R.J.P.M. Scholten

Dr. J.B. Reitsma

Overige leden: Prof. dr. J.A. Knottnerus

Dr. B.W.J. Mol

Prof. dr. M. Offringa

Prof. dr. P. Speelman

Prof. dr. T. Stijnen

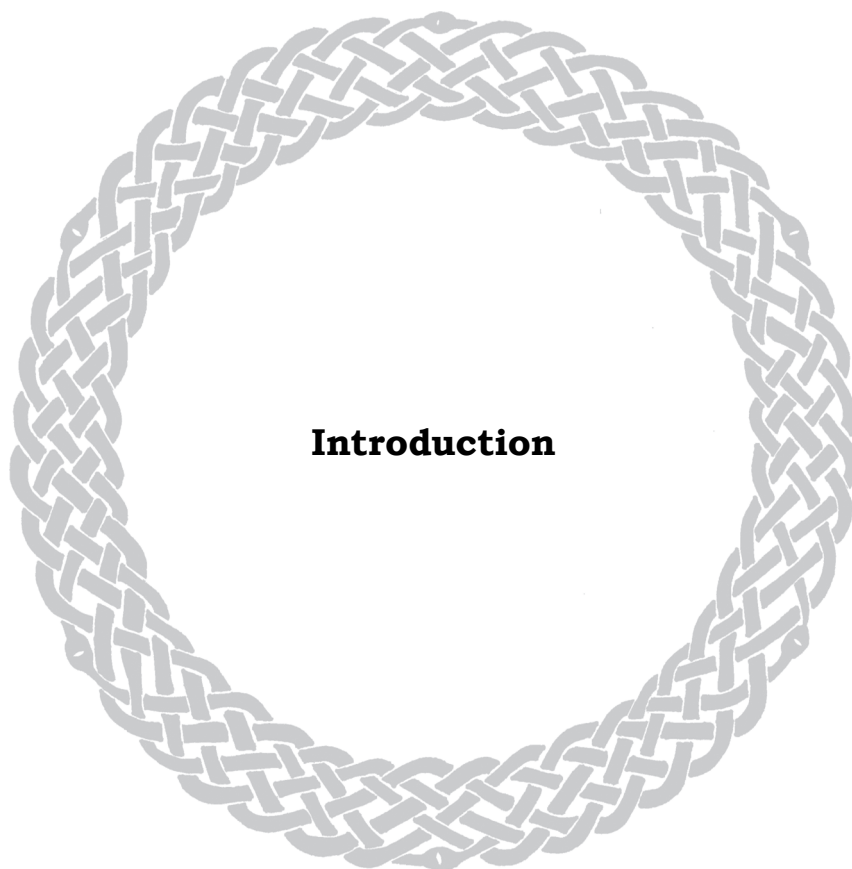
Prof. dr. H.C.W. de Vet

Faculteit der Geneeskunde

Table of Contents

Introduction	6
Chapter 1: Systematic Reviews of Diagnostic Test Accuracy – New Developments within The Cochrane Collaboration <i>Submitted for publication</i>	13
Chapter 2: The use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies <i>J Clin Epidemiol. 2006;59(3):234-40</i>	33
Chapter 3: Impact of adjustment for quality on results of meta-analyses of diagnostic accuracy <i>Clin Chem. 2007;53(2):164-72</i>	49
Chapter 4: Bias in sensitivity and specificity caused by data driven selection of optimal cut-off values: mechanisms, magnitude and solutions <i>Clin Chem. 2008; 54(4):729-37</i>	69
Chapter 5: Diagnostic accuracy may vary with prevalence: Implications for evidence-based diagnosis <i>Accepted by J Clin Epidemiol</i>	85
Chapter 6: Accuracy of fibronectin tests for the prediction of pre-eclampsia: a systematic review <i>Eur J Obstet Gynecol Reprod Biol. 2007;133(1):12-9</i>	101
Chapter 7: Galactomannan detection for the diagnosis of invasive aspergillosis in immunocompromized patients. A Cochrane Review of Diagnostic Test Accuracy <i>Conducted as a pilot Cochrane Diagnostic Test Accuracy review</i>	117
Chapter 8: Summary and discussion	161
Nederlandse samenvatting	169
Dankwoord	177
List of co-authors	183
List of publications	
Curriculum Vitae	187





Introduction

General introduction

Diagnostic tests aim to reduce uncertainty about an individual's condition. A plethora of tests is available for almost every condition imaginable. Examples include physical examination to rule out ankle fractures, mammograms to screen for breast cancer, magnetic resonance imaging for detecting herniated discs, portable chemical tests for blood glucose monitoring, nucleic acid amplification assays to detect infectious agents, and over the counter pregnancy tests.

A perfect test would identify all patients with the target condition, without making mistakes. This target condition may refer to a disease, or a disease stage, such as, for example, the healing phase of a fracture. Because perfect tests rarely exist, the users of a test may wish to know how well the test discriminates between individuals who have the target condition and those who have not. This is called diagnostic test accuracy.

The accuracy of a diagnostic test is studied by comparing the results of the test (or tests) under evaluation (also called index test) with the results of a reference standard. The reference standard is regarded as the best available method to establish the presence or absence of the target condition. The participants of a diagnostic accuracy study ideally undergo both the index test and the reference standard after which the results of both tests are compared (see Table 1). With dichotomous tests and a single target condition diagnostic test accuracy is often expressed as the proportion of people with the target condition who have indeed a positive test result (the test's sensitivity, or true positive fraction) and the proportion of people without the target condition who have a negative test result (the test's specificity, or true negative fraction)¹.

Table 1. 2 by 2 table

		Reference standard		
		+	-	
Index test	+	TP	FP	TP+FP
	-	FN	TN	FN+TN
		TP+FN	FP+TN	

In a 2 by 2 table, the results of the index test are compared with the results of the reference standard. TP=true positive; FP=false positive; FN=false negative; TN=true negative. Sensitivity = $TP/(TP+FN)$; Specificity = $TN/(TN+FP)$.

Central theme of the thesis

Health care professionals who are looking for evidence about how good a diagnostic test is in discriminating between patients with and without the target condition of interest, rely increasingly on systematic reviews of diagnostic test accuracy studies. Systematic reviews examine whether scientific findings are consistent and can be generalised across populations, settings, and treatment variations, or whether findings vary significantly by particular subsets^{2,3}.

Like any other research, the methodology of systematic reviews should be transparent and explicit, in order to minimise bias and maximise the informativeness in all parts of the review process. The methodology of systematic reviews involves the following steps

- (1) formulating a research question;
- (2) searching for the available evidence regarding the research question;
- (3) assessing the quality of the available evidence;
- (4) analysing the data;
- (5) interpreting the results.

The objective of this thesis is to provide empirical evidence to improve and guide the further development of the methodology behind systematic reviews of diagnostic test accuracy. The focus is specifically on the search process, incorporation of study quality, and analysis of the data.

Outline of the thesis

Chapter 1 gives an overview of the development of the methodology for diagnostic test accuracy systematic reviews over the last decade. The steps involved in a review are introduced. The following chapters offer a more detailed discussion of some of the features of a systematic review diagnostic test accuracy.

In **Chapter 2** we look at the usefulness of search strategies for retrieving diagnostic test accuracy studies in electronic bibliographical databases. We present the fraction of relevant studies that will be missed if search filters are used and we determine whether the search filters decrease the number of articles that one needs to screen to find one relevant article.

After retrieving the studies that are relevant for the systematic review, the quality of these studies needs to be assessed. The results of this quality assessment can be incorporated in the meta-analysis in many ways. In **Chapter 3** we compare three different strategies for incorporating quality, to test the hypothesis that adjustment for quality produces less optimistic estimates of diagnostic accuracy and narrower confidence intervals.

Chapter 4 addresses a possible source of bias when evaluating a test that produces a continuous result: post-hoc determination of an optimal cut-off value. We aim to determine the magnitude of bias in sensitivity and specificity associated with data-driven selection of cut-off values and to examine potential solutions to reduce this bias.

Chapter 5 addresses a possible source of heterogeneity between studies: differences in the prevalence of the target condition across studies. Although it is sometimes claimed that sensitivity and specificity do not depend on disease prevalence, we provide a number of real life examples in which accuracy varied with prevalence.

Chapters 6 and 7 are examples of systematic reviews of diagnostic test accuracy. In **Chapter 6** we report a review of the accuracy of fibronectin tests for the prediction of pre-eclampsia, one of the most important causes of maternal and fetal mortality and morbidity worldwide.

Chapter 7 presents a systematic review about the diagnostic accuracy of a commercially available galactomannan test for the diagnosis of invasive aspergillosis in immunocompromized patients. This systematic review served as a pilot review for the Cochrane Diagnostic Test Accuracy Working Group that was constituted to develop and test methods for the inclusion of diagnostic test accuracy reviews in The Cochrane Library⁴.

Chapter 8 summarizes the main findings of the research presented in this thesis and discusses a number of options for future research.

References

1. Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Serv Res.* 2002;2(1):4.
2. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med.* 1994;120(8):667-76.
3. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ.* 2001;323(7305):157-62.
4. The Cochrane Library, issue 1, 2008. <http://www3.interscience.wiley.com/cgi-bin/mr-whome/106568753/HOME?CRETRY=1&SRETRY=0>. Accessed 25 February 2008.





**Systematic Reviews of Diagnostic Test Accuracy
– New Developments within The Cochrane
Collaboration**

**Mariska M.G. Leeflang, Jon J. Deeks,
Constantine Gatsonis, Patrick M.M. Bossuyt**

Submitted for publication

Abstract

During the last two decades, the number of systematic reviews and meta-analyses of diagnostic test accuracy has grown considerably and substantial progress has been made in developing and agreeing on methodological standards.

The Cochrane Collaboration now considers it timely to register systematic reviews of diagnostic test accuracy studies, with the first Cochrane Diagnostic Test Accuracy Reviews scheduled to be published in the Cochrane Library in October 2008. Adding such reviews to the Cochrane Library may increase its appeal as the best single source of reliable evidence about the effects of health care.

Systematic review of diagnostic test accuracy studies can be methodologically challenging. Diagnostic accuracy studies can be difficult to identify. They are likely to show substantial variability, because of small sample sizes, clinical diversity, due to differences in setting or spectrum, and because of differences in design. Unlike randomized trials, which report a single measure such as the relative risk, diagnostic test accuracy studies usually report a pair of measures of test performance, such as the test's sensitivity and specificity, either at a point or along a ROC curve. Methods for meta-analysis have to take this bivariate nature of the data into account.

In this paper we present some of the most recent developments in the methodology for conducting systematic reviews and meta-analyses of diagnostic test accuracy studies that will be incorporated in the Cochrane review process.

1.1 Introduction

Diagnostic tests are a critical component of health care. Clinicians, policy makers and patients routinely face a range of questions regarding diagnostic tests. They want to know if testing improves outcome, would like to know what test to use, to purchase, or to recommend in practice guidelines, and how to interpret the results of testing.

Systematic reviews can help practitioners and decision-makers in answering these questions, by summarizing the available evidence and helping to explain differences among studies on the same question. The number of systematic reviews and meta-analyses of diagnostic test accuracy has grown remarkably in recent years. A search in MEDLINE (see Appendix 1.1) identified 77 published diagnostic reviews in 1996, a number that increased to 591 in 2006.

The Cochrane Collaboration is the largest international organization preparing, maintaining and promoting systematic reviews to help people make well-informed decisions about health care. In 2008 the 1st Issue of the Cochrane Database of Systematic Reviews (CDSR) included 3,384 reviews¹. Up until now, CDSR has been restricted to reviews of interventions, but the growing interest and the methodological advances in the synthesis of studies of diagnostic tests has lead to a change of policy, and from October 2008 CDSR will also include systematic reviews of diagnostic test accuracy.

The Cochrane Diagnostic Test Accuracy Working Group was launched in 2003 to systematize the approach and develop the software for these new systematic reviews. A meeting of more than 40 methodologists and expert reviewers from around the world was held in 2004 which reached consensus on appropriate methods and a reporting structure for protocols and reviews. Smaller working groups were subsequently formed to address each of the stages involved in the systematic review process. In the following years, these smaller groups reviewed methods and developed detailed guidance for review authors and review groups, which will be made available in the Cochrane Handbook for Diagnostic Test Accuracy Reviews². The methods in the Handbook are based on empirical evidence where available, making it a valuable resource for all authors of systematic reviews and meta-analyses of diagnostic test accuracy, including those preparing such reviews outside the scope of The Cochrane Collaboration.

In 1994, Irwig and colleagues presented guidelines for meta-analyses evaluating diagnostic tests in this journal³. We review the key methodological developments concerning problem formulation, location of literature, quality assessment and analysis that have occurred since then, using our experience from the work on the Handbook.

1.1.1 Diagnostic Test Accuracy Reviews

A study of the diagnostic accuracy of a test is undertaken to estimate the ability of that test to distinguish between patients with disease (or more generally, a specified target condition) and those without. In such a study, the results of the test under evaluation, or 'index test', are compared with those of the clinical reference standard determined in the same patients. The clinical reference standard is the best available method for classifying patients as having the target condition or not. Test accuracy is most often expressed as the test's sensitivity (the proportion of those positive to the reference standard who are also positive to the index test) and specificity (the proportion of those negative to the reference standard who are also negative to the index test), but many alternative measures have been proposed and are in use. The diagnostic accuracy of several tests may be evaluated in parallel in a single study.

Accuracy measures estimate the ability of a test to distinguish between persons with and without the target condition. Transformed to likelihood ratios, they can also be used to convert estimates of pre-test probabilities of disease to post-test probabilities, using Bayes' theorem. When a new test is supposed to replace an existing one, one has to find out how the accuracy of that test compares to the existing one⁴⁻⁶. More generally, accuracy can help clinicians to make decisions about tests and their future role⁷. Good accuracy is a desirable but not a sufficient condition for the effectiveness of that test. To show that using a new test does more good than harm in terms of patient outcomes, one may require randomized trials of test-and-treatment strategies.

As elsewhere in science, systematic reviews and meta-analyses can be used to obtain more precise estimates, when small studies addressing the same test and patients in the same setting are summarized. Systematic reviews can also be useful to establish whether and how scientific findings vary significantly by particular subsets, providing summary or subgroup estimates of diagnostic test accuracy that may be more applicable than estimates from a single study. They may help in identifying studies with the lowest risk of bias and they can be used to explore the between study heterogeneity in results. Such heterogeneity is to be expected, and probably even more so with diagnostic accuracy studies. Some of the variability is due to chance, as many diagnostic studies have small sample sizes⁸. Some will be due to differences in study methods, but study populations are also likely to differ between studies, resulting in differences in accuracy estimates⁹. Systematic reviews may also be used to address questions that were not directly considered in the primary studies, such as comparisons between tests.

In what follows, we briefly discuss the steps for conducting a systematic review of test accuracy studies (see Table 1.1). The account is our summary of the methods profiled in the Handbook for Cochrane diagnostic test accuracy reviews.

Table 1.1: The steps that are involved in systematic reviews of diagnostic test accuracy

- | | |
|----|---|
| 1. | Definition of the objectives of the review. |
| 2. | Identification of studies. |
| 3. | Quality assessment and applicability to the clinical problem at hand. |
| 4. | Data-analysis and presentation of the results. |
| 5. | Interpretation of the results. |

1.2 Definition of the objectives of the review

Any diagnostic research question should start with a precise description of the test or tests of interest, the disease or condition which they have to help identify, and a definition of the clinical context in which they will be used. From these statements inclusion criteria can be developed that define the studies of relevance to include in the review. A typical question is whether the test of interest has sufficient accuracy, in a well defined patient population, setting and testing strategy, to fulfil a particular role. Many such questions will be comparative, contrasting the accuracy of two or more tests or testing strategies.

The role of the test under evaluation relative to the current best practice needs to be specified, including its relative position to other tests used for the same target condition. Possible questions for a new test are: (1) can this test replace another test; (2) can it serve as a triage instrument, guiding further testing, and (3) can the test be used in addition to current best practice to pick up additional cases of the target disease, or to identify and eliminate false positives⁷. If a new test is to replace an existing test, then comparing the accuracy of both tests on the same population and with the same reference standard provides the most direct evidence. In the case of triage, one will be looking for a test that gives a minimal proportion of false negatives, so that the test can rule out disease in patients who will need no further testing. If the new test is to be used in addition to existing strategies, its aim will mainly be to reduce the number of false negatives, or, alternatively, the number of false positives. The review should provide data to assess the incremental change in accuracy made by adding the new test.

Test accuracy is not a fixed property of a test. It varies with the group of patients tested, with their spectrum of disease, with the clinical setting, with the test interpreters, and depends on the level of prior testing. For this reason, it is essential to include these elements in the study question.

1.2.1 Framing the research question

In a systematic review of the diagnostic accuracy of urinary markers for bladder cancer, the following issues were considered while defining the research question and objectives of the review¹⁰. In clinical practice, cytology was used to triage pa-

tients before they underwent invasive cystoscopy. As cytology combines a high specificity with a low sensitivity, the goal of the review was to identify a tumour marker with sufficient accuracy to either replace cytology or to be used in addition to cytology. For a marker to replace cytology, it has to combine an equally high specificity with a sufficiently high sensitivity, around 100%. From these objectives followed the inclusion criteria for the review. To include a study, markers and cytology had to be evaluated against the same reference standard, cystoscopy or histopathology; data to calculate sensitivity and specificity had to be available. Bladder tumours secondary to a cancer already identified and other target conditions were not allowed, as the diagnostic accuracy obtained in these cases cannot be translated directly to primary bladder tumours.

1.3 Identification of studies

Searching for and identifying test accuracy studies is now known to be more difficult than searching for randomized trials¹¹. There is not a clear, unequivocal key word or indexing term for an accuracy study, comparable to the term “randomized controlled trial”. The term “sensitivity and specificity” may look suitable, but is inconsistently applied in most databases. Data on diagnostic test accuracy may also be hidden in studies that did not have test accuracy estimation as their primary objective. This complicates the efficient identification of diagnostic test accuracy studies in electronic databases, such as MEDLINE. So until indexing systems are changed to properly code studies of test accuracy, searching for them will remain challenging.

In the development of a comprehensive search strategy, search strings that refer to the (1) test(s) under evaluation, (2) the target condition, (3) the patient description, or a subset of these can be used. For tests with a clear name that are used for a single purpose, just searching for publications in which those tests are mentioned may suffice. For other reviews, adding the patient description may be necessary, although this is also poorly indexed. A search strategy in MEDLINE should contain both MeSH headings and text words. If one searches for articles about tests for bladder cancer, for example, it will be necessary to include as many synonyms for bladder cancer as possible in the search strategy, including neoplasm, carcinoma, transitional cell and, possibly, also haematuria.

Several methodological electronic search filters for diagnostic test accuracy studies have been developed, which attempt to restrict a topic search to articles most likely to be test accuracy studies¹¹⁻¹⁴. These filters rely on indexing terms for research methodology and text words used in reporting results. However, they often miss relevant studies and are unlikely to decrease the number of articles one needs to screen, so are not recommended for use in systematic reviews^{15,16}. The incremental

effects of searching in languages other than English and in the so called grey literature have not yet been fully investigated.

In systematic reviews of intervention studies, publication bias is an important and well-studied form of bias. For clinical trials, the magnitude and determinants of publication bias have been identified by tracing the publication history of cohorts of trials reviewed by ethics committees and research boards. A consistent observation has been that studies with statistically significant results are more likely to be published than studies with non-significant findings. Investigating publication bias for diagnostic tests is problematic, as many studies are undertaken without ethical review or study registration, so follow-up of cohorts of studies is not possible¹⁷. Tests used in reviews of randomized controlled trials to detect publication bias have proven to be seriously misleading for diagnostic studies, and alternatives have poor power¹⁸. The determinants for publication of diagnostic studies may not be the same as the determinants for publication of intervention studies.

1.4 Assessment of methodological quality

Test accuracy studies with design deficiencies can produce biased results¹⁹⁻²¹. Sources of bias in test accuracy studies for which there is unambiguous evidence that diagnostic accuracy can be overestimated are the inclusion of healthy controls and the incomplete or differential use of reference standards^{19,21}. Quality assessment of individual studies in systematic reviews is therefore necessary to identify potential sources of bias and to limit the effects of these biases on the estimates and the conclusions of the review.

Based on the available evidence, Whiting and colleagues have used a Delphi procedure to develop the QUADAS checklist, now the recommended tool for quality assessment in diagnostic test accuracy studies²². The items that are listed in QUADAS relate to patient spectrum issues, verification issues, information bias and incomplete reporting. Issues related to the availability and quality of the reference standard include the overall appropriateness of this standard, partial or differential verification, important time gap between index test and reference standard, and the inclusion of the index test results in the reference standard in case of a composite reference standard.

The magnitude and direction of the resulting bias caused by methodological shortcomings may vary, depending on target condition and clinical setting. In addition, other items can be a cause of bias for specific tests. For example, in studies assessing the accuracy of biochemical serum markers, data-driven selection of the cut-off value may bias diagnostic accuracy^{23,24}. Review authors should therefore think carefully whether items need to be added to the QUADAS list.

Unfortunately, any evaluation of study quality is hampered by incomplete reporting²⁵. Guidelines for complete and transparent reporting have been developed²⁶, but their effects are only gradually becoming visible in the literature²⁷.

The results of quality appraisal can and should be summarized, to offer a general impression of the validity of the available evidence. Using an overall quality score is not recommended, as different shortcomings may generate different magnitudes of bias, even in opposing directions, making it very hard to attach sensible weights to each quality item²⁸. A way to summarize the quality assessment is shown in Figure 1.1, where stacked bars are used for each QUADAS item.

In the analysis phase, the results of the quality appraisal can be used to guide explorations of the sources of heterogeneity^{30,31}. Possible methods to address quality differences are sensitivity analysis, subgroup analysis or regression analysis, although the number of included studies may often be too low for meaningful meta-regression. The interpretation of results should at least be made bearing in mind the risk of bias.

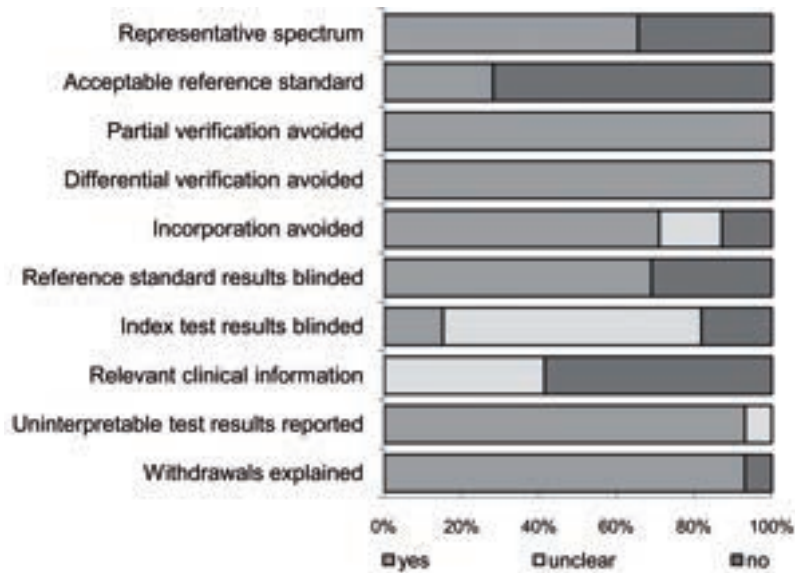


Figure 1.1. Review authors' judgments about quality items presented as percentages across all included studies.

Based on a re-analysis of data from a systematic review on magnetic resonance imaging for multiple sclerosis²⁹. The item "acceptable delay between tests" did not apply in this review. The authors considered the relative lack of acceptable reference standard as the main weakness of the review.

1.5 Analyzing the data and presenting the results

Whereas the results of a randomized trial are often reported using a single measure of effect, such as a difference in means or a risk difference or ratio, the results of most diagnostic test accuracy studies are reported with two or more statistics, the sensitivity and the specificity, the positive and negative predictive value, or likelihood ratios for the respective test results, or the ROC curve and quantities based on it^{32,33}.

The first step in the meta-analysis of diagnostic test accuracy is to visually examine the results of the individual studies. The paired results for sensitivity and specificity in the included studies can be plotted in a paired forest plot (see Figure 1.2) or plotted as points in an ROC plot (see Figure 1.3).

Plots of estimated sensitivity and specificity often display a pattern of negative correlation with each other across studies of the same test. A major contributor to this appearance is the trade-off between the true sensitivity and specificity of a test, as the threshold for defining test positivity varies. Decreasing the threshold that defines a test as positive rather than negative will increase sensitivity and decrease specificity (or *vice versa*), as described by the ROC curve for that test. When studies included in a review vary in positivity thresholds, a ROC-curve like pattern may be discerned across the points on the summary ROC plot.

There may be explicit variation in thresholds if different studies use different numerical thresholds to define a test result as positive (for example, variation in the blood glucose level above which a patient is said to have diabetes). In other situations, unquantifiable or implicit variation in threshold may occur when test results depend on interpretation or judgment (for example, between radiographers classifying images as normal or abnormal) or where test results are sensitive to machine calibration.

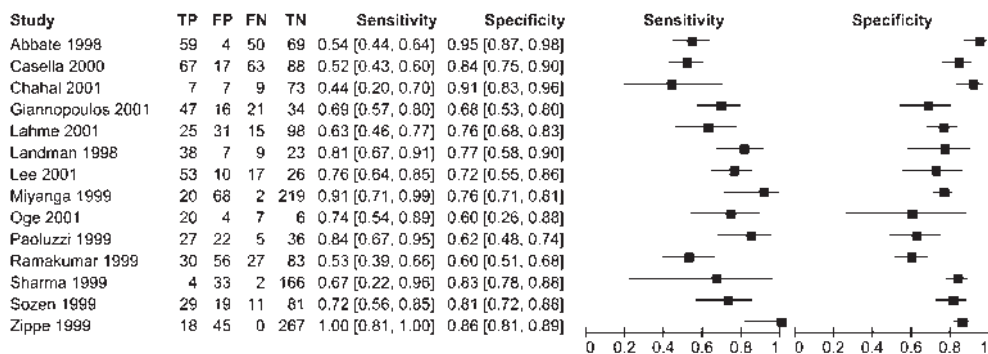


Figure 1.2. Forest plots of sensitivity and specificity of a tumor marker for bladder cancer. Based on a re-analysis of the data from Glas et al.¹⁰.

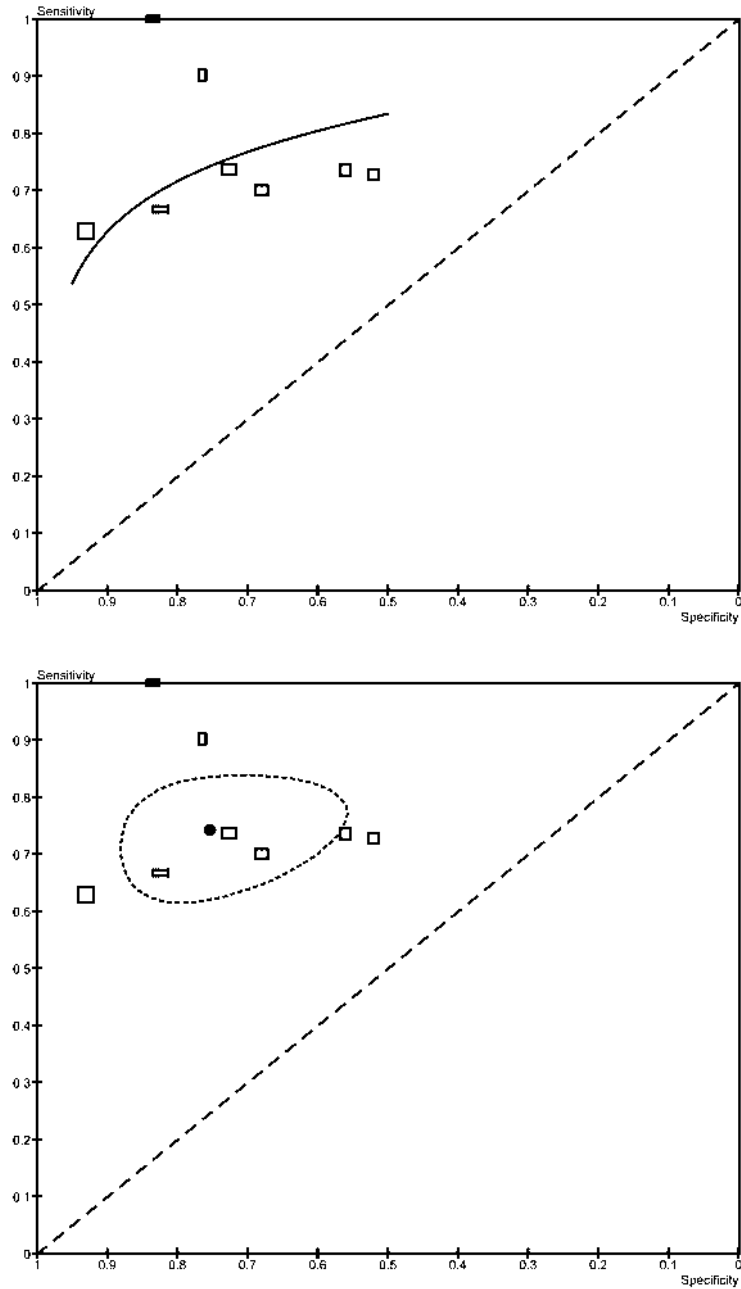


Figure 1.3a and b. ROC showing pairs of sensitivity and specificity values for the included studies. The height of the rectangles is proportional to the number of patients with bladder cancer across studies, the width of the rectangles corresponds to the number of patients without bladder cancer. Figure 1.3a shows the summary ROC curve that can be drawn through these values. Figure 1.3b shows the summary point estimate (black spot) and its 95% confidence region around it. Based on a re-analysis of the data from Glas et al.¹⁰.

Because threshold effects cause sensitivity and specificity estimates to appear as negatively correlated according to a ROC curve shape, and because threshold variation can be assumed to be present in nearly all situations to some degree, robust approaches to meta-analysis estimate the underlying relationship between sensitivity and specificity by constructing a summary ROC (SROC) curve. An average 'operating point' on this curve may subsequently be identified that indicates where the centre of the study results lie. Separate pooling of sensitivity and specificity to identify this point has been discredited, because in such an approach may identify a point which does not lie on the SROC curve when there is between study variation.

In 1994, Irwig and colleagues³ recommended Moses and Littenberg's linear regression model for the construction of summary ROC curves³⁴, which is based on regressing the log diagnostic odds ratio against a measure of the proportion reported as test positive. Extending the regression model by adding covariates has been proposed to examine differences between tests and relate them to study or sample characteristics. However, the formulation of the model has limitations in failing to consider the precision of the study estimates, not estimating between study heterogeneity and the explanatory variable in the regression being measured with error. These problems render estimates of confidence intervals and *P*-values that are unsuitable for formal inference^{33,35}.

Two approaches to fitting random effects hierarchical models have been developed to overcome these limitations: the hierarchical summary ROC (HSROC) model^{33,36,37} and the bivariate random effects model^{35,38}. The HSROC model focuses on identifying the underlying ROC curve, estimating the average accuracy and average threshold (and unexplained variation in these parameters across studies), together with a shape parameter that describes the asymmetry in the curve. The bivariate random effects model focuses on estimating the average sensitivity and specificity, but also estimates the unexplained variation in sensitivity and specificity and the correlation between them. These two basic models have been shown to be mathematically equivalent. Both can be used to identify the underlying SROC curve and the average operating point^{35,39}. They can also be used to explore heterogeneity by adding covariates to the models, or by applying separate models to different subgroups. Both models can be fitted with statistical software for fitting mixed models^{33,35,37,38}.

Some authors have advocated the pooling of likelihood ratios rather than pooling of sensitivity and specificity or pooling of ROC curves^{40,41}. However, summary likelihood ratios can be easily calculated with the methods described above, while calculating sensitivity and specificity from pooled likelihood ratios may result in sensitivities and specificities above 1 or below 0⁴².

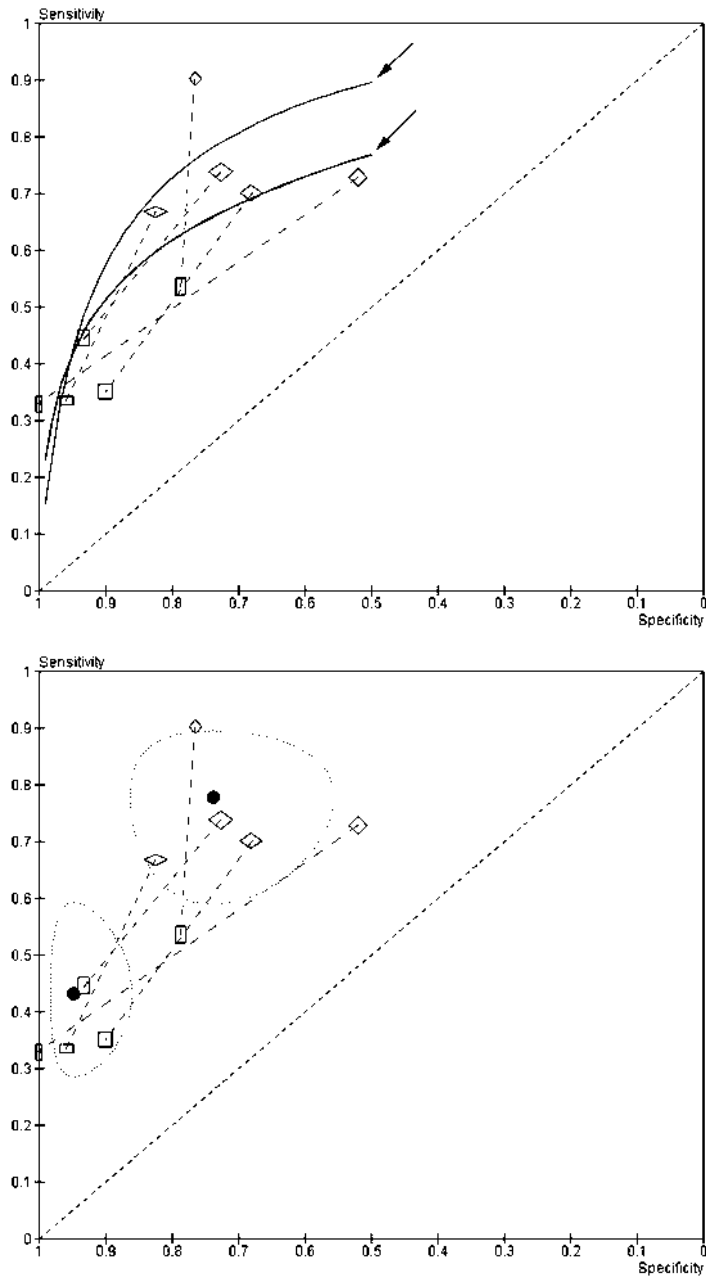


Figure 1.4. Direct comparison of two index tests for bladder cancer: cytology (squares) and bladder tumor antigen (diamonds).
 Figure 1.4a shows the summary ROC curve that can be drawn through these values. Figure 1.4b shows the summary point estimate of sensitivity and specificity (black spot) and its 95% confidence region around it. The two tests clearly show a trade-off between sensitivity and specificity: cytology has a significantly higher specificity (ellipse closest to Y-axis lower arrow points at ROC curve) and BTA has a significantly higher sensitivity (higher ellipse and arrow points at highest ROC curve). It will depend on the role of the test in practice which test is considered 'best'. Based on a re-analysis of the data from Glas et al.¹⁰.

1.5.1 Curves or summary points

The ability to estimate both underlying SROC curves and average operating points allows flexibility in testing hypotheses and estimating diagnostic accuracy. Analyses to estimate underlying SROC curves can be based on all included studies, and facilitate well powered comparisons between different tests, or between subgroups of studies, which are not restricted to investigating accuracy at a particular threshold. An example can be found in Figure 1.3a, where the diagnostic accuracy of a bladder tumour antigen test for diagnosing bladder cancer is summarized with an SROC curve. In contrast, estimation of a summary point specific to a test being used at a common threshold is useful to obtain the best estimate of test accuracy in parameters clinicians understand. The certainty associated with the estimate can be described by confidence regions marked on the SROC plot around the average point. An example of this approach is given in Figure 1.3b.

1.5.2 Comparative analyses

Systematic reviews of diagnostic test accuracy may evaluate more than one test to determine which test or combination of tests can better serve the intended purpose. Indirect comparisons can be made by calculating separate summary estimates of the sensitivity and specificity of each test including all studies that have evaluated that test, regardless of whether they evaluated the other tests. The substantial variability that can be expected between tests means that such comparisons are prone to confounding. Restricting inclusion to studies of similar design and patient characteristics may limit the confounding.

An alternative approach is to only use studies that directly compared the tests in the same patients, or randomized patients to tests. Such direct comparisons do not suffer from confounding. Unfortunately, fully paired studies are not always available. Paired analyses can be displayed in an ROC plot, by linking the sensitivity-specificity pairs from each study with a dashed or dotted line, as in Figure 1.4.

1.6 Interpretation of the results

The interpretation of the results offered in the systematic review should help readers to understand the implications for practice. This interpretation should consider whether evidence derived from the review is actually suitable for addressing the objectives of the review, and not consist solely of reporting the results. The interpretation of the findings should consider the consequences of the false positive and false negative test results and whether the estimates of accuracy that were found are sufficiently high for the foreseen role that the test will have in practice. A decision model could be used to structure the interpretation of the findings. Such a model would incorporate important factors as the disease prevalence and the available diagnostic and therapeutic interventions that may follow the test.

Some reviews may not result in useful summary estimates of sensitivity and specificity, for example because of large variability in the individual study estimates, or because the authors only investigated the comparative accuracy by comparing SROC curves. The potential effects of quality differences, or the lack of high quality studies on the results should be considered. Additional information, such as costs or important trade-offs between harms and benefits can be included.

1.7 Conclusion

Important progress has been made in recent years in the methods for developing methodology for systematic reviews of diagnostic test accuracy studies. We know more about searching, about sources of bias in study design, and about quality appraisal. In meta-analysis new hierarchical random effects models have been developed with sound statistical properties. Methods for the estimation of summary ROC curves and of summary estimates of sensitivity and specificity are now available. All these advances are described in detail in the Cochrane Handbook for Diagnostic Test Accuracy Reviews.

Diagnostic test accuracy reviews face two major challenges. Firstly, they are limited by the quality and availability of primary test accuracy studies that address important relevant questions. More studies are needed which recruit a suitable spectrum of participants, make direct comparisons between tests, use rigorous methodology, and clearly report their methods and findings. Secondly, more development is needed in the area of interpretation and presentation of the results of diagnostic test accuracy reviews. It has been shown that many clinicians struggle with the definitions of sensitivity, specificity and likelihood ratios^{41,42}. Furthermore, policy makers and guideline developers may be interested in the comparative accuracy only, or in additional information, such as costs and burden. Developing systematic reviews that are really relevant for both policy makers and clinical practice poses a major challenge, and clear thinking about the scope and purpose of the review is a necessary condition.

The Cochrane Diagnostic Test Accuracy Working Group addresses those challenges and will continue developing, evaluating and disseminating the methods for diagnostic test accuracy reviews.

Contributors to the Cochrane Diagnostic Test Accuracy Working Group include (in alphabetical order): Bert Aertgeerts, Doug Altman, Gerd Antes, Lucas Bachmann, Patrick Bossuyt, Heiner Buchner, Peter Bunting, Frank Buntinx, Jonathan Craig, Jon Deeks, Jenny Doust, Matthias Egger, Anne Eisinga, Constantine Gatsonis, Paul Glasziou, Roger Harbord, Jorgen Hilden, Lotty Hooft, Andrea Horvath, Chris Hyde, Les Irwig, Monica Kjeldstrøm, Petra Macaskill, Susan Mallett, Ruth Mitchell, Tess Moore, Rasmus Moustgaard, Wytze Oosterhuis, Madhukar Pai, Prashni Paliwal, Daniel Pewsner, Hans Reitsma, Jacob Riis, Ingrid Riphagen, Anne Rutjes, Rob Scholten, Nynke Smidt, Jonathan Sterne, Yemisi Takwongi, Riekje de Vet, Vasivy Vlassov, Joseph Watine, Danielle van der Windt, Penny Whiting.

References

1. The Cochrane Library, issue 1, 2008. http://www.mrw.interscience.wiley.com/cochrane/cochrane_clsystev_articles_fs.html. Accessed 23 January 2008.
2. Cochrane Diagnostic Test Accuracy Working Group. <http://srdta.cochrane.org/en/index.html>. Accessed 25 February, 2008.
3. Irwig L, Tosteson AN, Gatsonis CA, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. *Ann. Int. Med.* 1994; 120(8):667–676.
4. Thornbury JR. Clinical efficacy of diagnostic imaging: love it or leave it. *AJR.* 1994;162(1):1–8.
5. Knottnerus JA. (Ed) *The Evidence Base of Clinical Diagnosis*. BMJ Books, London, 2002.
6. 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(11):850–5.
7. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ.* 2006;332(7549):1089–92.
8. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ.* 2006;332(7550):1127–9.
9. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ.* 2002;324(7338):669–71.
10. Glas AS, Roos D, Deutekom M, Zwinderman AH, Bossuyt PM, Kurth KH. Tumor markers in the diagnosis of primary bladder cancer. A systematic review. *J Urol.* 2003;169(6):1975–82.
11. Haynes RB, Wilczynski N, McKibbin KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Inform Assoc.* 1994;1(6):447–58.
12. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol.* 2000;53(1):65–9
13. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *JAMA.* 2002;9(6):653–8.
14. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from MEDLINE: analytical survey. *BMJ.* 2004;328(7447):1040.
15. Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. *J Clin Epidemiol.* 2005;58(5):444–9.
16. Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *J Clin Epidemiol.* 2006;59(3):234–40.
17. Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol.* 2002;31(1):88–95.
18. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol.* 2005;58(9):882–93.
19. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 1999 Sep 15;282(11):1061–6.
20. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med.* 2004;140(3):189–202.

21. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*. 2006;174(4):469–76.
22. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003;3:25
23. Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in Sensitivity and Specificity Caused by Data-Driven Selection of Optimal Cutoff Values: Mechanisms, Magnitude, and Solutions. *Clin Chem*. 2008;54(4):729–37.
24. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. *J Clin Epidemiol*. 2006;59(8):798–801.
25. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, Bouter LM, de Vet HC. Quality of reporting of diagnostic accuracy studies. *Radiology*. 2005;235(2):347–53.
26. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med*. 2003;138(1):40–4.
27. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, Bouter LM, de Vet HC. The quality of diagnostic accuracy studies since the STARD statement: has it improved? *Neurology*. 2006;67(5):792–7.
28. Whiting P, Harbord R, Main C, Deeks JJ, Filippini G, Egger M, Sterne JA. Accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis: systematic review. *BMJ*. 2006;332(7546):875–84.
29. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol*. 2005;5:19.
30. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? *BMC Med Res Methodol*. 2005;5(1):20.
31. Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deeks J, Bossuyt P. Impact of adjustment for quality on results of meta-analyses of diagnostic accuracy. *Clin Chem*. 2007;53(2):164–72.
32. Zhou X-H, Obuchowski N, McClish D. *Statistical methods in diagnostic medicine*. Wiley, 2002.
33. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. *AJR Am J Roentgenol*. 2006;187(2):271–81.
34. Moses LE, Littenberg B, Shapiro D. Combining Independent Studies of a Diagnostic Test Into a Summary ROC Curve: Data-Analytic Approaches and Some Additional Considerations. *Stat Med*. 1993;12(14):1293–1316.
35. Arends LR. *Multivariate meta-analysis: modeling the heterogeneity. Mixing apples and oranges: dangerous or delicious? Haveka BV, Alblasserdam, 2006*
36. Rutter C. and Gatsonis C. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20(19):2865–84.
37. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol*. 2004;57(9):925–32
38. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol*. 2005;58(10):982–90.
39. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics*. 2007;8(2):239–51.
40. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics*. 2003;59(4):936–46.

Chapter 1

41. Khan KS. Systematic reviews of diagnostic tests: a guide to methods and application. *Best Pract Res Clin Obstet Gynaecol.* 2005;19(1):37-46.
42. Khan KS, Dinnes J, Kleijnen J. Systematic reviews to evaluate diagnostic tests. *Eur J Obstet Gynecol Reprod Biol.* 2001;95(1):6-11.
43. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Stat Med.* 2008;27(5):687-97
44. Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. *BMJ.* 2002;324(7341):824-6.
45. Puhan MA, Steurer J, Bachmann LM, ter Riet G. A randomized trial of ways to describe test accuracy: the effect on physicians' post-test probability estimates. *Ann Intern Med.* 2005;143(3):184-9.

Appendix 1.1. Search for diagnostic reviews

PubMed search strategy for identification of diagnostic test accuracy reviews in MEDLINE:

("Diagnosis"[Majr] OR diagnos*[ti] OR accuracy[ti]) AND (meta-analysis[tw] OR systematic review[tw]).

PubMed accessed on 5th March, 2008.





**The use of methodological search filters
to identify diagnostic accuracy studies can
lead to the omission of relevant studies**

**Mariska M.G. Leeflang, Rob J.P.M. Scholten,
Anne W.S. Rutjes, Johannes B. Reitsma, Patrick M.M. Bossuyt**

J Clin Epidemiol. 2006;59(3):234-40

Abstract

Objective: to determine the usefulness of methodological filters in search strategies for diagnostic studies in systematic reviews.

Methods: we made an inventory of existing methodological search filters for diagnostic accuracy studies and applied them in PubMed to a reference set derived from 27 published systematic reviews in a broad range of clinical fields. Outcome measures were the fraction of not identified relevant studies and the reduction in the number of studies to read.

Results: we tested twelve search filters. Two to 28% of the studies included in the systematic reviews did not pass the sensitive search filters, 4 to 24% did not pass the accurate filters and 39% or 42% did not pass the specific filters. Decrease in Number Needed to Read when a search filter was used in a search strategy for a diagnostic systematic review varied from 0% to 77%.

Conclusion: the use of methodological filters to identify diagnostic accuracy studies can lead to the omission of a considerable number of relevant studies, otherwise included. When preparing a systematic review, it may be preferable to refrain from the use of methodological filters.

2.1 Introduction

Systematic reviews are regarded as the cornerstones of evidence based medicine. They aim to identify and evaluate all available evidence about a specific topic. A systematic and comprehensive search for relevant primary studies is one of the essential steps in conducting a systematic review, and one of the factors that distinguishes a systematic review from a traditional narrative review.

To identify diagnostic accuracy studies in an electronic database, such as MEDLINE, several search strategies can be used. Many of these strategies rely on free text words and MeSH headings directed to disease indicators in combination with search terms for the diagnostic test. To further limit the search results, a methodological filter can be used consisting of text words and MeSH headings directed to general indicators of diagnostic studies. However, in contrast to intervention studies, these general indicators are not widely and systematically used as keywords for diagnostic studies and indexing of original diagnostic accuracy studies is not flawless¹⁻³. Diagnostic studies show more variability in study design than intervention studies. It is therefore not unlikely that a considerable number of relevant studies will be missed when those filters are used in diagnostic reviews while the reduction in number of studies to consider for inclusion is far from impressive.

A number of these methodological search filters for diagnostic accuracy studies has been validated. They are known to differ in sensitivity (percentage correctly identified studies) and specificity (percentage correctly non-identified studies). The aim of this study was to assess the usefulness of these search filters by applying them to a reference set that was derived from published systematic reviews in a broad range of clinical fields. First, the fraction of relevant studies that did not pass each filter was calculated. Then we determined whether the diagnostic search filters focus the search strategy enough to be practical.

2.2 Methods

To identify articles reporting on the development of diagnostic search filters, we performed a computerized search using the databases MEDLINE, EMBASE and the Cochrane Methodology Register of the Cochrane Library, all until January 2004. The search terms used in MEDLINE (interface PubMed) were: "(MEDLINE[MeSH] OR "Information Storage and Retrieval/methods"[MeSH]) AND diagnosis". The search terms for EMBASE (interface OVID) were ((search adj strategy).mp. or (search adj strategies).mp.) and (diagnosis.mp or exp DIAGNOSIS/), where "Exp" means exploded and '/' stands for Emtree term. The Cochrane Methodology Register was searched with "diagnosis" only. Two reviewers independently assessed papers for inclusion. Any disagreement was resolved by consensus. Papers were included if one of the main objectives was the development and validation of a diagnostic search strategy

to be used in MEDLINE. We used no language restriction. Of each included study we selected the filters with the highest sensitivity (proportion of relevant articles that correctly passed the filter), highest accuracy (the highest possible sensitivity in combination with the highest possible specificity) and highest specificity (proportion irrelevant articles that correctly did not pass the filter), according to the authors, for further evaluation. Most of these filters were developed and tested using the OVID interface. As PubMed is the only freely available search engine for MEDLINE and the most widely used, we converted the OVID-search terms into PubMed-format.

The converted search filters were applied to a reference set consisting of studies that had been included in systematic reviews of diagnostic test accuracy. These reviews were selected after an electronic search for systematic reviews of diagnostic accuracy studies published between January 1999 and April 2002 in MEDLINE, EMBASE, the Database of Abstracts of Reviews of Effect (DARE) and the MEDION database of the University of Maastricht (Figure 2.1). This search strategy is available from the authors. Criteria for inclusion were the assessment of diagnostic accuracy, the inclusion of more than 10 original studies with inclusion not based on design characteristics, and sufficient data to reproduce the contingency table. We excluded those systematic reviews that reported the application of diagnostic search filter.

For each review and for all reviews combined we calculated the fraction of primary studies that would have been missed by each filter. This fraction was also calculated per year of publication of the original studies. In addition, we selected

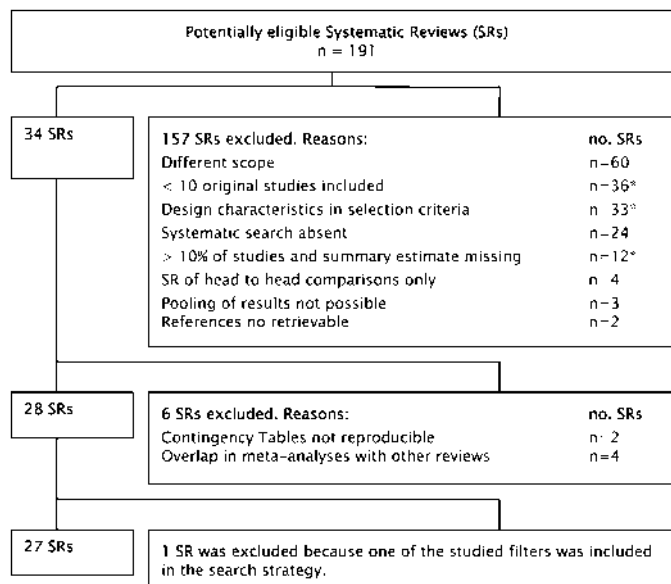


Figure 2.1. Inclusion of the systematic reviews.

On the right side, the excluded reviews are listed with reasons for exclusion. * = some exclusion criteria can overlap; SR = Systematic Review.

the systematic reviews with a clearly reported search strategy. We replicated these searches in PubMed using the same time frame as the authors used. The results of these searches were combined with the sensitive and accurate filters to find out to what extent the filters reduced the number needed to read (i.e. the number of articles needed to read to identify one relevant article, calculated as 1 divided by precision)⁴.

2.3 Results

Our search strategy for diagnostic search filters revealed a total number of 781 articles from three databases (MEDLINE 605, EMBASE 95 and Cochrane Methodology Register 81). Eight articles were included⁴⁻¹¹. These articles described a total of 28 validated diagnostic search filters. Two studies described each one accurate filter^{7,8} and these filters were both selected as accurate filters. One study did not report sensitivity, specificity, nor accuracy of filters and the filters described here were a basic filter and an extended filter⁵. We selected the extended filter as most sensitive one. One study only reported sensitivities, so we selected the most sensitive filter only¹⁰. Two other studies reported explicitly sensitivities and specificities, so we selected the most sensitive and specific filter from each study^{4,9}. Two studies reported sensitivities, specificities and accuracy^{5,11}. The thus selected six sensitive, two specific and four accurate filters were then applied to the reference set of articles that were included in the systematic reviews (Table 2.1). The filters of Bachmann⁴ and Haynes and colleagues^{5,11} were the only filters not specifically developed for use in systematic reviews. Nine selected filters consisted of MeSH terms as well as free words or text words. These filters all contained the MeSH term "Sensitivity and Specificity". One of the filters of Devillé only consisted of free words. The specific filter of Haynes and Wilczynski only contained the text word 'specificity'¹¹.

The 27 selected systematic reviews varied in target disease of interest and in index test studied (Table 2.2): seven regarding laboratory diagnosis, five regarding physical examination, 15 regarding diagnostic imaging and one regarding history taking. They included 11 to 110 original articles, with a total of 921 articles¹²⁻³⁸. Of these, 29 were not stored in MEDLINE, eight of which could be retrieved in EMBASE, four were unpublished and 17 were retrieved via various other pathways. Of the remaining 892 studies, 72 were used in two or more reviews, leaving 820 individual original articles as reference set in our study.

Table 2.2 and Figure 2.2 show the performance of the selected filters, based on the identification percentages as achieved for the reference list of each review. On average, 14% of the studies included in a systematic review did not pass the sensitive search filters (range 0 to 92% for the individual filter-review combinations), 13% did not pass the accurate filters (range 0 to 87% for the individual filter-review combinations) and 41% did not pass the specific filters (range 0 to 100%). There was a

Table 2.1. Details of the 12 identified filters, transcribed to PubMed format.

Filter ^a	Search terms
H94se ⁵	"sensitivity and specificity"[MeSH] OR diagnosis[subheading:noexp] OR "diagnostic use"[subheading] OR sensitivity[tw] OR specificity[tw]
H94sp ⁵	"sensitivity and specificity"[MeSH] OR (predictive[tw] AND value[tw])
H94acc ⁵	"sensitivity and specificity"[MeSH] OR "Diagnosis"[MeSH] OR "diagnostic use"[subheading] OR specificity[tw] OR (predictive[tw] AND value[tw])
VDW97 ⁶	"Diagnosis"[MeSH] OR "sensitivity and specificity"[MeSH] OR "Reference Values"[MeSH] OR "False Positive Reactions"[MeSH] OR "False Negative Reactions"[MeSH] OR "Mass Screening"[MeSH] OR diagnos* OR sensitivity OR specificity OR predictive value* OR reference value* OR ROC* OR likelihood ratio* OR monitoring
D00acc ⁹	"sensitivity and specificity"[MeSH] OR specificity[tw] OR false negative[tw] OR accuracy[tw]
D00se ⁹	"sensitivity and specificity"[MeSH] OR specificity[tw] OR false negative[tw] OR screening[tw]
B02 ⁴	"Sensitivity and Specificity"[MeSH] OR predict* OR diagnose* OR diagnosi* OR diagnost* OR accura*
V03 ¹⁰	"sensitivity and specificity"[MeSH] OR sensitivity[tw] OR specificity[tw] OR predictive value*[tw] OR false positiv*[tw] OR false negativ*[tw] OR observer variation*[tw] OR roc curve*[tw] OR likelihood ratio*[tw] OR "Likelihood Functions"[MeSH]
D02a ⁷	specificity OR screening OR false positive OR false negative OR accuracy OR (predictive AND value*) OR (reference value*) OR ROC OR likelihood ratio
D02b ⁸	"Sensitivity and Specificity"[MeSH] OR "mass screening"[MeSH] OR "Reference values"[MeSH] OR specificit*[tw] OR screening[tw] OR false positive*[tw] OR false negative*[tw] OR accuracy[tw] OR predictive value*[tw] OR reference value*[tw] OR roc*[tw] OR likelihood ratio*[tw]
H04se ¹¹	sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnos*[Title/Abstract] OR diagnosis[MeSH:noexp] OR diagnostic * [MeSH:noexp] OR diagnosis,differential[MeSH:noexp] OR diagnosis[Subheading:noexp]
H04sp ¹¹	specificity[tw]

^a Label used in this study and reference to source

large difference between reviews. Some articles from reference lists easily passed all sensitive filters, whereas from other reference lists the filters blocked all of the reference articles, even if they were all stored in MEDLINE. The systematic review in which filters behaved most problematically was the one about gallstones. The articles included in this review were only indexed with terms referring to population (adult, women) and surgery. Of the 23 articles included in the review, eleven were indexed with "/diagnosis" and only one was indexed with "sensitivity and specificity". The diagnostic filter as described by Van der Weijden et al.⁶ had the lowest percentages of incorrectly withheld articles (2%), whereas the sensitive filter developed by Vincent¹⁰

Table 2.2. Proportion of original articles missed by each filter per review and all reviews combined.

Systematic Review	Searched from:	MEDLINE searched:	Target Disease	No. of refs. in MEDLINE	sensitive filters						accurate filters						specific filters	
					H945e	VDM97	D004e	B02	V03	H045e	H94acc	D00acc	D02a	D02b	H94sp	H04sp		
Balk, 2001	1966	Yes	Cardiac ischemia	45	0.00	0.00	0.16	0.00	0.13	0.00	0.07	0.18	0.00	0.16	0.29	0.29		
Berger, 2000	1966	Yes	Gallstones	23	0.43	0.30	0.74	0.30	0.74	0.30	0.43	0.87	0.35	0.70	0.78	0.96		
Deville, 2000	1992	Yes	Herniated discs	33	0.31	0.08	0.92	0.08	0.85	0.08	0.46	0.85	0.31	0.77	0.85	1.00		
Fiellin, 2000	1966	Yes	Alcohol problems	38	0.03	0.00	0.11	0.00	0.24	0.00	0.11	0.26	0.00	0.11	0.46	0.37		
Gould, 2001	1966	Yes	Pulmonary lesions	40	0.03	0.00	0.03	0.00	0.03	0.05	0.00	0.03	0.00	0.03	0.15	0.08		
Hobby, 2000	1985	Yes	Wrist problems	15	0.07	0.00	0.47	0.07	0.40	0.07	0.00	0.33	0.00	0.33	0.67	0.53		
Hoffman, 2000	1986	Yes	Prostate cancer	21	0.05	0.00	0.05	0.00	0.05	0.00	0.00	0.05	0.00	0.05	0.05	0.14		
Hoogendam, 1999	1983	Yes	Prostate cancer	13	0.23	0.00	0.00	0.08	0.23	0.08	0.00	0.38	0.00	0.00	0.38	0.62		
Huicho, 2002	1966	Yes	Urinary tract infection in children	46	0.15	0.04	0.28	0.07	0.37	0.07	0.17	0.50	0.13	0.26	0.50	0.57		
Hurlley, 2000	1966	Yes	Endotoxemia	51	0.49	0.00	0.82	0.41	0.75	0.00	0.45	0.00	0.00	0.00	0.90	0.88		
Kelly, 2001	1975	Yes	Gastro-oesophageal carcinoma	24	0.42	0.00	0.58	0.04	0.54	0.25	0.00	0.29	0.00	0.29	0.63	0.67		
Kim, 2001	1975	Yes	Coronary disease	80	0.00	0.00	0.11	0.00	0.10	0.04	0.00	0.09	0.00	0.09	0.24	0.15		
Koelmay, 2001	1985	Yes	Lower extremity arterial disease	34	0.00	0.00	0.12	0.00	0.12	0.00	0.00	0.12	0.00	0.12	0.15	0.18		
Kwok, 1999	1966	Yes	Coronary disease	23	0.00	0.00	0.13	0.00	0.09	0.04	0.00	0.04	0.00	0.04	0.30	0.26		
Lau, 2001	1966	Yes	Cardiac ischemia	110	0.15	0.04	0.37	0.14	0.35	0.14	0.06	0.32	0.12	0.30	0.45	0.46		
Ledrick, 1999	1966	Yes	Abdominal aortic aneurysms	42	0.17	0.05	0.64	0.12	0.81	0.10	0.17	0.18	0.05	0.00	0.88	0.00		
LI, 2001	1966	No	Emergency intubation	10	0.2	0.00	0.20	0.20	0.10	0.20	0.00	0.20	0.20	0.20	0.40	0.30		
Minchell, 1999	1966	Yes	Cervical cancer	37	0.11	0.00	0.16	0.05	0.38	0.08	0.00	0.32	0.00	0.16	0.51	0.68		
Mol, 1999	1990	Yes	Fetus with Down syndrome	25	0.12	0.00	0.12	0.12	0.20	0.12	0.00	0.00	0.00	0.04	0.40	0.40		
Nellemans, 2000	1991	Yes	Peripheral arterial disease	21	0.00	0.00	0.10	0.00	0.10	0.00	0.00	0.10	0.00	0.10	0.10	0.19		
Safriel, 2000	1990	Yes	Pulmonary emboli	10	0.10	0.00	0.10	0.00	0.10	0.00	0.00	0.10	0.00	0.10	0.10	0.10		
Sloan, 2000	1985	Yes	Sexually transmitted diseases	25	0.04	0.00	0.08	0.00	0.04	0.00	0.08	0.16	0.00	0.00	0.16	0.16		
Smith-Bindmann, 2001	1980	Yes	Fetus with Down syndrome	56	0.13	0.00	0.16	0.11	0.23	0.13	0.00	0.38	0.04	0.16	0.41	0.50		
Sonnud, 2001	1984	Yes	Prostate cancer	22	0.05	0.00	0.27	0.00	0.14	0.14	0.00	0.14	0.00	0.05	0.36	0.27		
Vasquez, 2000	1984	NR	Acute cholecystitis	17	0.06	0.00	0.35	0.00	0.24	0.06	0.06	0.35	0.00	0.29	0.59	0.47		
Visser, 2000	1984	Yes	Peripheral arterial disease	27	0.00	0.00	0.04	0.00	0.04	0.00	0.00	0.04	0.00	0.04	0.11	0.07		
Westwood, 2002	1990	Yes	Carotid stenosis	24	0.04	0.00	0.25	0.00	0.29	0.00	0.00	0.17	0.00	0.13	0.42	0.38		
			All 27 reviews	820	0.19	0.08	0.54	0.12	0.56	0.13	0.19	0.59	0.12	0.49	0.71	0.72		
			All 27 reviews	mean	0.12	0.02	0.27	0.07	0.28	0.07	0.08	0.24	0.04	0.17	0.42	0.39		
			All 27 reviews	median	0.07	0.00	0.16	0.00	0.23	0.06	0.00	0.18	0.00	0.11	0.40	0.37		

NR = not reported

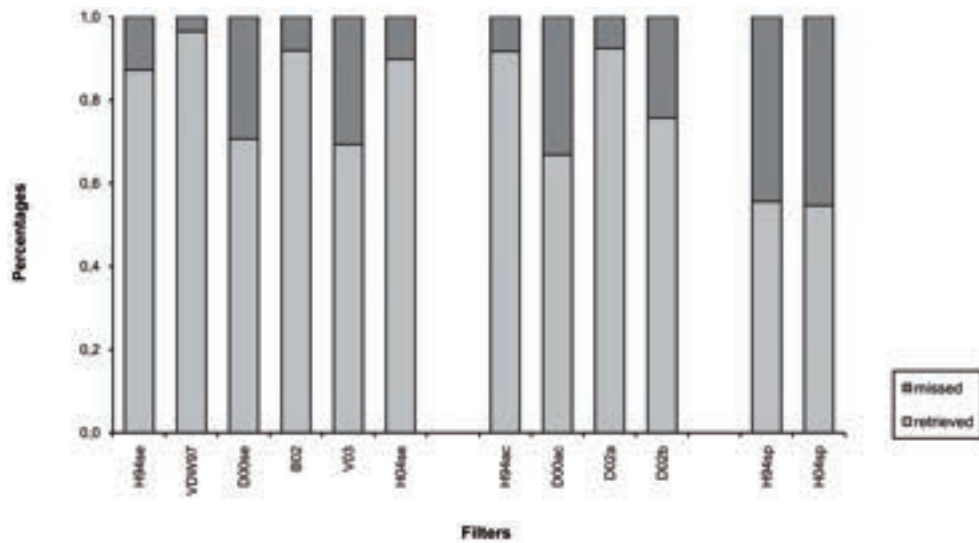


Figure 2.2. Average proportion of retrieved and missed references per filter.

Average proportion of retrieved and missed references per filter, based on the identification percentages as achieved for the reference list of each review. For the filters, see Table 2.1.

caused the highest loss in articles (28%) of all the sensitive filters. The accurate filters had the same percentages of articles that did and did not pass the filters as the sensitive ones. There was no relationship between number of search terms or usage of certain search terms and number of articles missed by the sensitive and accurate filters. When plotted against year of publication, the average percentage of primary studies that will be missed by a search filter decreases (Figure 2.3). This effect was even clearer for the specific filters and not seen for the filter of Van der Weijden et al..⁶ (results not shown).

The last step in this study was to determine the reduction in the number of studies needed to read to identify one relevant article after applying each search filter. Of only six reviews, the search strategy was reported in such a way, that it was possible to replicate the search undertaken by the authors of the review. In these six studies the period in which the original search was conducted was also reported well (month and year). The combinations of these strategies with the sensitive and accurate search filters lead to a reduction in Number Needed to Read of 0 to 169 studies, representing a relative decrease ranging from 0% to 77% (Table 2.3). The search filters with the largest decrease in Number Needed to Read also had the highest number of missed articles.

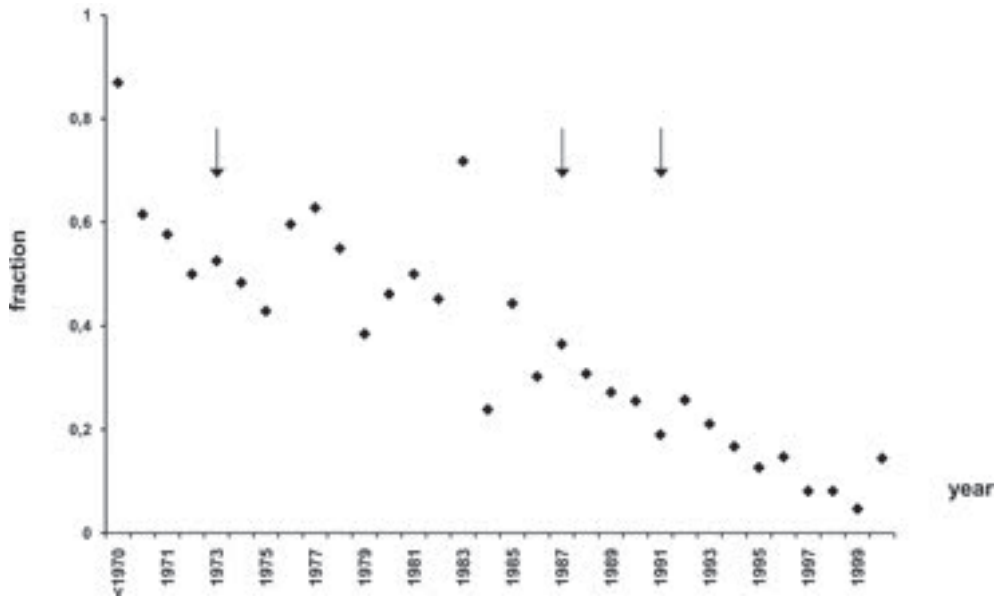


Figure 2.3. Proportion of not identified articles per year, the average of all sensitive filters. The arrows indicate the year that some MeSH terms are introduced: diagnosis (both as MeSH terms and as subheading) and diagnostic use (subheading) were introduced before 1970; False Negative Reactions[MeSH] in 1973, Predictive Value[MeSH] in 1987 and Sensitivity and Specificity[MeSH] in 1991.

Table 2.3. Reduction of Number Needed to Read for six reviews by applying methodological filters.

	Hobby			Hoffman			Kelly			Koelemay			Kwok			Mol		
	Total [†]	NNR	Ex*	Total [†]	NNR	Ex*	Total [†]	NNR	Ex*	Total [†]	NNR	Ex*	Total [†]	NNR	Ex*	Total [†]	NNR	Ex*
No Filter	241	24.1	5	1523	72.5	0	4874	211.9	1	7008	219.0	2	5505	275.3	3	1830	76.3	1
VDW97	240	24.0	5	1412	67.2	0	4687	203.8	1	6916	216.1	2	5505	275.3	3	1431	59.6	1
D02a	234	23.4	5	1251	59.6	0	4426	192.4	1	5987	187.1	2	5505	275.3	3	1297	54.0	1
B02	199	19.9	5	1119	53.3	0	3582	162.8	2	5731	179.1	2	5250	262.5	3	1009	45.9	3
H04se	182	18.2	5	943	44.9	0	3321	184.5	6	5282	165.1	2	4549	239.4	4	945	43.0	3
H94acc	239	23.9	5	1312	62.5	0	4468	194.3	1	6792	212.3	2	5505	275.3	3	1278	53.3	1
H94se	169	16.9	5	823	41.2	1	2343	167.4	10	5143	160.7	2	5171	258.6	3	671	30.5	3
D02b	112	14.0	7	711	35.6	1	1646	102.9	8	1802	62.1	5	2008	105.7	4	629	27.3	2
D00acc	104	13.0	7	577	28.9	1	1509	94.3	8	1465	50.5	5	1727	90.9	4	272	19.4	11
D00se	94	15.7	9	661	33.1	1	1259	125.9	14	1467	50.6	5	1659	97.6	6	566	27.0	4
V03	100	16.7	9	608	30.4	1	1326	120.5	13	1453	50.1	5	1926	107.0	5	385	20.3	6

[†]= the total number of studies left after applying the strategy and the filter. NNR = the Number of articles Needed to Read to find one relevant article. *= The number of relevant studies that were missed using this strategy and this filter.

2.4 Discussion

Diagnostic reviews aim to identify and evaluate all available evidence about a specific index test or a comparison of tests. If the yield of the initial search based on index test and target condition is too large, a diagnostic search filter could be helpful to reduce this number. In this study we compared the performance of 12 validated diagnostic search filters by applying them to a set of articles, selected from the reference lists of 27 published diagnostic systematic reviews. All filters studied may lead to the non-inclusion of relevant studies, varying from an average of 2% of the total number of relevant primary articles used in this study to 42%. Only one of the reviews present in the initial set of 28 reviews to be used in this study reported the use of a methodological search filter for diagnostic studies and seven reviewers did not mention any search term at all. If these seven reviews used one of the search filters reported here, the results are overestimated and the real percentage of missed studies can be even higher.

When performing a systematic review of diagnostic accuracy studies, the application of methodological filters may lead to incomplete retrieval of evidence. This effect is even aggravated in studies published before 1990. Vincent and colleagues have pointed out that the various published search filters had a lower sensitivity when applied to a set of articles with a broader range of publication data compared to a hand searched reference set¹⁰. Haynes and colleagues have compared search strategies for the years 1991 and 1986 and they also concluded that these filters performed less in the earlier published studies⁵. After adapting the search strategies to the 1986 database by using MeSH terms that were in use to indicate diagnostic studies at that time, the poorer performance could not be improved. These effects are probably caused by a poorer indexing in earlier years, especially before 1986. Furthermore, some MeSH terms have only been indexed since recent years. For example, the MeSH term “sensitivity and specificity” was introduced in 1991, whereas the subheading “/diagnosis” was already in use in 1966³⁹.

As proposed by Haynes and colleagues⁵ in 1994, search filters can be regarded as diagnostic test for identification of articles that can be included in a systematic review. Due to lack of information about falsely identified positively and falsely not identified articles, we were not able to construct two by two tables, as is customary in diagnostic accuracy studies. We used the reference lists of systematic reviews as reference set for the filters, but we have no information about whether the authors of the reviews had adequate literature searches. Our aim was to check whether the methodological search filters were able to detect at least those publications that were included in the reviews. Articles that might have been missed by the original review, can either be retrieved or missed by the methodological search filters we evaluated, leading to an under- or over-evaluation of the search filter respectively. We have not formally assessed the impact of incompleteness of the reference set. We however reason that the effect will be minor, since the methodological filters did not function substantially different in reviews that had used elaborate search strate-

gies compared to reviews that used less elaborate search strategies. In contrast with other validation studies, our set covers a broader range of journals, publication years and clinical topics. Furthermore, we did not assess the methodological quality of the primary studies that were missed by the filters. Search filters that have been designed for retrieval of methodologically strong studies^{5,11} may therefore perform worse in our set of primary articles than in the original set, where the filter was developed. Search filters also vary strongly in their percentage of missed studies per review and per target condition. However, we have no data on the effects that these missing studies would have had on the summary outcome estimate. This would be a worthwhile objective for further research.

Rewriting a search filter from OVID-format (or other formats) to PubMed-format is another factor that may influence the performance of the filters (number of missed studies and decrease in Number Needed to Read). The use of acronyms, search tags and search fields is not the same in both interfaces⁴⁰. We have tested the most sensitive filter of Haynes and Wilczynski¹¹ as it is used in Clinical Queries of PubMed⁴¹. In a response to Haynes' article¹¹, Falck-Ytter and Motschall presented another way to transcribe the original OVID-query to PubMed-format⁴². The results of this search term differ from the results of the Clinical Queries search term. Although it is not yet clear if any of these uncertainties lead to a different set of included articles in a review - and, possibly, other conclusions - reviewers have to keep these comments in mind when they are about to use search filters in PubMed.

We did not only study the number of studies that would be missed by using the search filters, we also looked whether the use of a diagnostic search filter would reduce the Number Needed to Read sufficiently to be practically useful. The use of these filters did not always lead to a substantial decrease in total number of articles identified by combining search terms for just patient group and test. The filters that led to a greater reduction did so at the expense of missing more relevant studies. We can not be completely sure if the search strategies were perfectly replicated, as the authors of the reviews did not always report the interface they used to search through MEDLINE. However, we think this study shows in general that search filters not always lead to the decrease in articles that is wished for. Furthermore, three of the studies from which we selected the search filters reported that their filters were developed for the use in clinical practice^{4,5,11}. These filters did not decrease the number needed to read more than the other filters.

The use of search filters for diagnostic studies inevitably leads to the loss of relevant articles. This is due in part to uncertainties about whether filters based on other interfaces perform equally after transcription in PubMed, but the major reasons are the poor indexing of diagnostic studies in MEDLINE and the wide range of possible designs for diagnostic accuracy studies. Because search filters are not guaranteed to reduce the number of studies, they cannot be expected to increase search efficiency. We think that the use of diagnostic search filters in the development of a systematic review should be discouraged. In practice however, the urge for retriev-

Chapter 2

ing all relevant studies might be less. And if no useful systematic reviews can be found, the filters may be helpful to lighten at least a part of the job.

References

1. Fielding AM, Powell A. Using MEDLINE to achieve an evidence-based approach to diagnostic clinical biochemistry. *Ann Clin Biochem* 2002; 39(Pt 4):345-350.
2. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ* 1994; 309(6964):1286-1291.
3. Wilczynski NL, Walker CJ, McKibbin KA, Haynes RB. Reasons for the loss of sensitivity and specificity of methodologic MeSH terms and textwords in MEDLINE. *Proc Annu Symp Comput Appl Med Care* 1995;436-440.
4. Bachmann LM, Coray R, Estermann P, ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *J Am Med Inform Assoc* 2002; 9(6):653-658.
5. Haynes RB, Wilczynski N, McKibbin KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Inform Assoc* 1994; 1(6):447-458.
6. van der Weijden T, IJzermans CJ, Dinant GJ, van Duijn NP, de Vet R, Buntinx F. Identifying relevant diagnostic studies in MEDLINE. The diagnostic value of the erythrocyte sedimentation rate (ESR) and dipstick as an example. *Fam Pract* 1997; 14(3):204-208.
7. Deville WL, Bossuyt PM, de Vet HC, Bezemer PD, Bouter LM, Assendelft WJ. [Systematic reviews in practice. X. Searching, selecting and the methodological assessment of diagnostic evaluation research]. *Ned Tijdschr Geneesk* 2002; 146(48):2281-2284.
8. Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol* 2002; 2(1):9.
9. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol* 2000; 53(1):65-69.
10. Vincent S, Greenley S, Beaven O. Clinical Evidence diagnosis: Developing a sensitive search strategy to retrieve diagnostic studies on deep vein thrombosis: a pragmatic approach. *Health Info Libr J* 2003; 20(3):150-159.
11. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from MEDLINE: analytical survey. *BMJ* 2004; 328(7447):1040.
12. Balk EM, Ioannidis JP, Salem D, Chew PW, Lau J. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: a meta-analysis. *Ann Emerg Med* 2001; 37(5):478-494.
13. Berger MY, van der Velden JJ, Lijmer JG, de Kort H, Prins A, Bohnen AM. Abdominal symptoms: do they predict gallstones? A systematic review. *Scand J Gastroenterol* 2000; 35(1):70-76.
14. Deville WL, van der Windt DA, Dzaferagic A, Bezemer PD, Bouter LM. The test of Lasegue: systematic review of the accuracy in diagnosing herniated discs. *Spine* 2000; 25(9):1140-1147.
15. Fiellin DA, Reid MC, O'Connor PG. Screening for alcohol problems in primary care: a systematic review. *Arch Intern Med* 2000; 160(13):1977-1989.
16. Gould MK, Maclean CC, Kuschner WG, Rydzak CE, Owens DK. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. *JAMA* 2001; 285(7):914-924.
17. Hobby JL, Tom BD, Bearcroft PW, Dixon AK. Magnetic resonance imaging of the wrist: diagnostic performance statistics. *Clin Radiol* 2001; 56(1):50-57.
18. Hoffman RM, Clanon DL, Littenberg B, Frank JJ, Peirce JC. Using the free-to-total prostate-specific antigen ratio to detect prostate cancer in men with nonspecific elevations of prostate-specific antigen levels. *J Gen Intern Med* 2000; 15(10):739-748.

19. Hoogendam A, Buntinx F, de Vet HC. The diagnostic value of digital rectal examination in primary care screening for prostate cancer: a meta-analysis. *Fam Pract* 1999; 16(6):621-626.
20. Huicho L, Campos-Sanchez M, Alamo C. Meta-analysis of urine screening tests for determining the risk of urinary tract infection in children. *Pediatr Infect Dis J* 2002; 21(1):1-11, 88.
21. Hurley JC. Concordance of endotoxemia with gram-negative bacteremia. A meta-analysis using receiver operating characteristic curves. *Arch Pathol Lab Med* 2000; 124(8):1157-1164.
22. Kelly S, Harris KM, Berry E, Hutton J, Roderick P, Cullingworth J et al. A systematic review of the staging performance of endoscopic ultrasound in gastro-oesophageal carcinoma. *Gut* 2001; 49(4):534-539.
23. Kim C, Kwok YS, Heagerty P, Redberg R. Pharmacologic stress testing for coronary disease diagnosis: A meta-analysis. *Am Heart J* 2001; 142(6):934-944.
24. Koelemay MJ, Lijmer JG, Stoker J, Legemate DA, Bossuyt PM. Magnetic resonance angiography for the evaluation of lower extremity arterial disease: a meta-analysis. *JAMA* 2001; 285(10):1338-1345.
25. Kwok Y, Kim C, Grady D, Segal M, Redberg R. Meta-analysis of exercise testing to detect coronary artery disease in women. *Am J Cardiol* 1999; 83(5):660-666.
26. Lau J, Ioannidis JP, Balk EM, Milch C, Terrin N, Chew PW et al. Diagnosing acute cardiac ischemia in the emergency department: a systematic review of the accuracy and clinical effect of current technologies. *Ann Emerg Med* 2001; 37(5):453-460.
27. Lederle FA, Simel DL. The rational clinical examination. Does this patient have abdominal aortic aneurysm? *JAMA* 1999; 281(1):77-82.
28. Li J. Capnography alone is imperfect for endotracheal tube placement confirmation during emergency intubation. *J Emerg Med* 2001; 20(3):223-229.
29. Mitchell MF, Cantor SB, Brookner C, Utzinger U, Schottenfeld D, Richards-Kortum R. Screening for squamous intraepithelial lesions with fluorescence spectroscopy. *Obstet Gynecol* 1999; 94(5 Pt 2):889-896.
30. Mol BW, Lijmer JG, van der MJ, Pajkrt E, Bilardo CM, Bossuyt PM. Effect of study design on the association between nuchal translucency measurement and Down syndrome. *Obstet Gynecol* 1999; 94(5 Pt 2):864-869.
31. Nelemans PJ, Leiner T, de Vet HC, van Engelshoven JM. Peripheral arterial disease: meta-analysis of the diagnostic performance of MR angiography. *Radiology* 2000; 217(1):105-114.
32. Safriel Y, Zinn H. CT pulmonary angiography in the detection of pulmonary emboli: a meta-analysis of sensitivities and specificities. *Clin Imaging* 2002; 26(2):101-105.
33. Sloan NL, Winikoff B, Haberland N, Coggins C, Elias C. Screening and syndromic approaches to identify gonorrhoea and chlamydial infection among women. *Stud Fam Plann* 2000; 31(1):55-68.
34. Smith-Bindman R, Hosmer W, Feldstein VA, Deeks JJ, Goldberg JD. Second-trimester ultrasound to detect fetuses with Down syndrome: a meta-analysis. *JAMA* 2001; 285(8):1044-1055.
35. Sonnad SS, Langlotz CP, Schwartz JS. Accuracy of MR imaging for staging prostate cancer: a meta-analysis to examine the effect of technologic change. *Acad Radiol* 2001; 8(2):149-157.
36. Vasquez TE, Rimkus DS, Hass MG, Larosa DI. Efficacy of morphine sulfate-augmented hepatobiliary imaging in acute cholecystitis. *J Nucl Med Technol* 2000; 28(3):153-155.
37. Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US--a meta-analysis. *Radiology* 2000; 216(1):67-77.

38. Westwood ME, Kelly S, Berry E, Bamford JM, Gough MJ, Airey CM et al. Use of magnetic resonance angiography to select candidates with recently symptomatic carotid stenosis for surgery: systematic review. *BMJ* 2002; 324(7331):198.
39. NLM. US National Library of Medicine. MeSH database. <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh>. Accessed 1 August 2004.
40. Gault LV, Shultz M, Davies KJ. Variations in Medical Subject Headings (MeSH) mapping: from the natural language of patron terms to the controlled vocabulary of mapped lists. *J Med Libr Assoc* 2002; 90(2):173-180.
41. NLM. US National Library of Medicine. Clinical Queries. <http://www.ncbi.nlm.nih.gov/entrez/query/static/clinicaltable.html> . Accessed 1 August 2004.
42. Falck-Ytter YT, Motschall E. New search filter for diagnostic studies: Ovid and PubMed versions not the same. <http://bmj.bmjournals.com/cgi/eletters/328/7447/1040#60819>. Accessed 1 June 2004.





**Impact of adjustment for quality on results
of meta-analyses of diagnostic accuracy**

**Mariska M.G. Leeflang, Johannes B. Reitsma,
Rob J.P.M. Scholten, Anne W.S. Rutjes, Marcello Di Nisio,
Jon J. Deeks, Patrick M.M. Bossuyt**

Clin Chem. 2007;53(2):164-72

Abstract

Background: We examined whether and to what extent different strategies of defining and incorporating quality of included studies affect the results of meta-analyses of diagnostic accuracy.

Methods: We evaluated the methodological quality of 487 diagnostic accuracy studies in 30 systematic reviews with the QUADAS (Quality Assessment of Diagnostic accuracy Studies) checklist. We applied three strategies that varied both in the definition of quality and in the statistical approach to incorporate the quality-assessment results into meta-analyses. We compared magnitudes of diagnostic odds ratios, widths of their confidence intervals, and changes in a hypothetical clinical decision between strategies.

Results: Following two definitions of quality, we concluded that only 70 or 72 of 487 studies were of “high quality.” This small number was partly due to poor reporting of quality items. None of the strategies for accounting for differences in quality led systematically to accuracy estimates that were less optimistic than ignoring quality in meta-analyses. Limiting the review to high-quality studies considerably reduced the number of studies in all reviews, with wider confidence intervals as a result. In 18 reviews, the quality adjustment would have resulted in a different decision about the usefulness of the test.

Conclusions: Although reporting the results of quality assessment of individual studies is necessary in systematic reviews, reader wariness is warranted regarding claims that differences in methodological quality have been accounted for. Obstacles for adjusting for quality in meta-analyses are poor reporting of design features and patient characteristics and the relatively low number of studies in most diagnostic reviews.

3.1 Introduction

Health care professionals seeking the best information about diagnostic tests increasingly turn to systematic reviews of test-accuracy studies, yet a review's summary estimate can be biased if the studies in the review are flawed. An evaluation of the quality of the original studies, therefore, is an essential issue of any systematic review.

The methodological quality of studies can be defined in terms of their susceptibility to bias. Studies with methodological shortcomings, such as inclusion of healthy control individuals or selective use of multiple reference standards to verify index test results, have produced different measures of test accuracy¹⁻⁵. In most cases, such deficiencies have been associated with inflated estimates of diagnostic accuracy. The inclusion of lower-quality studies in a meta-analysis may therefore produce unrealistically high accuracy estimates. Accounting for quality differences can be expected to produce less optimistic summary estimates of diagnostic accuracy.

Design feature variability and the presence of studies with sub optimal designs in a systematic review may also increase heterogeneity in results among studies⁶⁻⁸. Given these considerations, one can expect strategies that account for quality in meta-analyses of diagnostic accuracy to lead to more homogeneous results and therefore to more precise estimates, with narrower confidence intervals around the accuracy measures of interest, than estimates without quality adjustment.

Quality assessment of individual studies in a review may identify both design deficiencies that can lead to bias and sources of variation that can lead to heterogeneity. Several quality-assessment tools, most of which use a "checklist" approach, have been developed for diagnostic accuracy studies⁵. A recently developed generic quality-assessment tool based on a modified Delphi procedure^{5, 9} has been recommended by the Cochrane Collaboration as a starting point for quality assessment in diagnostic reviews¹⁰.

Although quality appraisal has been recognized as an essential step of systematic reviews, how study quality should be addressed in meta-analyses of diagnostic accuracy studies is less clear^{5,11}. Strategies to incorporate study quality into meta-analyses can be broadly divided into 3 categories: including all studies, irrespective of quality; analyzing subgroups that differ in quality; and multivariable regression analysis. The slightly different recommendations given in the guiding reports are all based on sparse evidence¹²⁻¹⁴.

To test the hypothesis that adjustment for quality produces less optimistic estimates of diagnostic accuracy and narrower confidence intervals, we compared 3 different strategies for incorporating quality in analyzing a number of previously published systematic reviews of diagnostic accuracy studies.

3.2 Methods

We studied 3 alternative strategies for quality incorporation in meta-analysis and comparing the results of analyzing all available studies irrespective of their quality, in a series of systematic reviews of diagnostic accuracy studies. Within each systematic review, we compared the summary diagnostic odds ratios (DORs) and the widths of the confidence intervals across these strategies.

3.2.1 Study set

To include a broad sample of diagnostic studies that examined a variety of tests over time, we conducted a systematic electronic search for systematic reviews of diagnostic accuracy studies published between January 1999 and April 2002⁵. This search produced a set of 28 reports of systematic reviews¹⁵⁻⁴². Details of the search strategy are available from the authors. Inclusion criteria were (1) a systematic review of diagnostic test-accuracy studies, (2) inclusion of at least 10 original studies, (3) no exclusion of primary studies based on design features, and (4) the ability to reproduce the 2 by 2 tables from the original studies. The 28 reports yielded 30 systematic reviews. Details of the inclusion process are reported elsewhere⁵.

A variety of conditions and index tests were studied in these 30 reviews (Table 3.1). The median number of studies in a review was 14 (interquartile range, 10 to 20). The median sample size of the individual studies was 100 (interquartile range, 43 to 288).

3.2.2 Assessment of methodological quality

We assessed the methodological quality of all 487 studies included in the 30 reviews with items from the QUADAS instrument⁹ (Table 3.2). We limited ourselves to the 7 QUADAS items most closely related to methodological quality and did not use the items that referred to quality of reporting. We dichotomized each item by scoring as deficient any study feature that was not reported.

QUADAS item 1 (Table 3.2) refers to both the generalizability of results and the possibility that the study may produce biased results. We assessed 3 patient-spectrum components that refer to the distorted selection of participants, because previous studies have linked these components to biased accuracy estimates. These components were consecutive enrollment of patients, case-control or 2-gate design vs. cohort design, and avoidance of limited challenge^{2,4}. Limited challenge was defined as the exclusion of patients with disease characteristics that may produce false-positive or false-negative results (e.g., exclusion of patients with existing lung disorders in an accuracy study of spiral computed tomography for the diagnosis of pulmonary embolism). A 2-gate study was defined as a case-control study in which cases and controls are sampled from 2 distinct source populations by means of different selection criteria⁴³.

Two independent assessors conducted quality assessments, and consensus meetings resolved disagreements. If necessary, a third person made the final decision.

Table 3.1. Characteristics of the systematic reviews in our study set.

Reference	Target condition	Index test(s)	No. of included studies
Balk et al., 2001 ¹⁵	Acute myocardial infarction	Laboratory test	9
Berger et al., 2000 ¹⁶	Gallstones	Physical examination	12
Deville et al., 2000 ¹⁷	Herniated discs	Physical examination	11
Fiellin et al., 2000 ¹⁸	Alcohol abuse	Questionnaires	14
Gould et al., 2001 ¹⁹	Pulmonary nodules	FDG-PET ^b	29
Hobby et al., 2000 ²⁰	Tears in wrist cartilage	MRI	11
Hoffman et al., 2000 ²¹	Prostate cancer	Laboratory test	21
Hoogendam et al., 1999 ²²	Prostate cancer	Physical examination	13
Huicho et al., 2002 ²³	Urinary tract infection	Laboratory test	18
Hurley, 2000 ²⁴	Gram-negative infections	Laboratory test	27
Kelly et al., 2001 ²⁵	Gastroesophageal carcinoma	Ultrasound	13
Kim et al., 2001 ²⁶	Coronary artery disease	Echocardiography	40
Koelemay et al., 2001 ²⁷	Peripheral arterial disease	MRA	9
Kwok et al., 1999 ²⁸	Coronary artery disease	Echocardiography	19
Lau et al., 2001 ²⁹	Acute myocardial infarction	Laboratory test	10
Lederle et al., 1999 ³⁰	Abdominal aortic aneurysm	Physical examination	10
Li, 2001 ³¹	Endotracheal tube placement	Capnography	10
Mitchell et al., 1999 ³²	Cervix lesions	Cytology	17
Mol et al., 1999 ³³	Down syndrome	Ultrasound	23
Nelemans et al., 2000 ³⁴	Peripheral arterial disease	MRA	13
Safriel et al., 2002 ³⁵	Pulmonary emboli	CT	10
Sloan et al., 2000 ³⁶	Gonorrhea and chlamydial infection	Physical examination	14
Smith-Bindman et al., 2001 ³⁷	Down syndrome	Ultrasound	28
Sonnad et al., 2001 ³⁸	Prostate cancer	MRI	21
Vasquez et al., 2000 ³⁹	Acute cholecystitis	Scintigraphy	15
Visser et al., 2000 ⁴⁰	Peripheral arterial stenosis	Ultrasound	17
Westwood et al., 2002 ⁴¹	Carotid stenosis	MRA	24
Wiese et al., 2000 ⁴²	Vaginal trichomoniasis	Cytology	29

^b FDG-PET, [¹⁸F]fluorodeoxyglucose positron emission tomography; MRI, magnetic resonance imaging; MRA, magnetic resonance angiography; CT, computed tomography.

Table 3.2. QUADAS items included in the 2 definitions of “high quality.”

	Evidence-based definition	Common-practice definition
1. Was the spectrum of patients representative of the patients who will receive the test in practice?		X
2. Were selection criteria clearly described?		
3. Is the reference standard likely to correctly classify the target condition?		
4. Is the time period between reference standard and index test short enough?		
5. Did the whole sample receive verification using a reference standard for diagnosis?	X	X
6. Did patients receive the same reference standard regardless of the index test results?	X	X
7. Was the reference standard independent from the index test?		
8. Was the execution of the index test described in sufficient detail to permit replication of the test?		
9. Was the execution of the reference standard described in sufficient detail to permit replication of the test?		
10. Were the index test results interpreted without knowledge of the results of the reference standard?	X	
11. Were the reference standard results interpreted without knowledge of the results of the index test?	X	
12. Were the same clinical data available when test results were interpreted as would be available in practice?		
13. Were uninterpretable / intermediate results reported?		
14. Were withdrawals from the study explained?		

3.2.3 Meta-analysis

We used the summary ROC model of Moses and Littenberg for our meta-analysis⁴⁴⁻⁴⁶. Their model uses linear regression analysis to examine how D, the natural logarithm of the DOR, changes as a function of S, which is the sum of $\text{logit}(\text{sensitivity})$ and $\text{logit}(1 - \text{specificity})$. S is related to the threshold for classifying a test as positive.

We modelled the intercept and slope of the model as fixed effects but included a random effect to allow for variation beyond chance among studies⁴⁷. We weighted studies by the inverse of the variance of the log DOR to allow for the precision with which each study measured the log DOR. This procedure gave more weight to larger studies.

In the multivariable quality-adjustment strategies, covariates representing quality items were added to the model; this step allowed the intercept and slope in the regression analysis to differ between subgroups of studies defined by the correspond-

ing covariate. In all strategies, we estimated the summary DOR over all studies the meta-analysis at the mean S value of these studies. Because the DOR cannot be calculated in 2 by 2 tables containing a zero, we added 0.5 to all 4 cells in these situations as a continuity correction^{44,48}.

3.2.4 Strategies for incorporating quality

We compared the following 3 statistical approaches to account for quality in meta-analyses: (1) The “restrict” strategy applied to meta-analysis of high-quality studies only. Studies were regarded as “high quality” when they fulfilled all quality criteria. (2) The “adjust all” strategy involved multivariable adjustment for all individual quality items by including all these items in a single multivariable model, irrespective of the strength of the association between these items and the DOR. (3) The “selective adjustment” strategy consisted of multivariable adjustment for only those quality items that were significantly associated with the DOR in a univariable analysis (P for entry <0.2)^{49,50}.

These strategies were compared with a reference strategy in which all studies within the original meta-analysis were included, irrespective of their quality characteristics.

Differences in results between strategies may depend both on the definition of quality and on the statistical approach used. We therefore considered 2 different sets of quality items to define higher-quality studies. The first set was chosen because there is empirical evidence that they can lead to biased results^{4,5}. This set, referred to as the “evidence-based” quality definition, includes QUADAS items 5, 6, 10, and 11 (Table 3.2). The second set of quality items (QUADAS items 1, 5, and 6) is referred to as the “common practice” quality definition and was selected because these 3 items are often applied in diagnostic reviews^{5,11}. The restrict strategy and the adjust-all strategy were applied twice, once with the evidence-based definition of quality and once with the common-practice definition.

3.2.5 Comparisons and analysis

We compared the summary DOR and its 95% confidence interval for the reference strategy, which included all studies, with the 3 quality-adjusting strategies in all 30 systematic reviews. Differences in results between strategies were analyzed within each systematic review with the Wilcoxon signed rank test to determine whether a strategy consistently led to higher or lower estimates of diagnostic accuracy. To investigate whether the strategies that adjusted for quality also resulted in more precise summary DOR estimates, we again used the Wilcoxon signed rank test statistic to compare the different approaches with respect to the absolute widths of the natural logarithm of the 95% confidence interval around the mean DOR.

To determine whether the change in summary DOR would affect clinical decisions, we used 4 arbitrary categories, which were defined by the absolute size of the summary DOR. If a meta-analysis resulted in a point estimate of the DOR <16 , the test

was regarded as not useful. We regarded a test with a DOR of 16 to 81 as moderately useful, a test with a DOR of 81 to 361 as useful, and a test with a DOR >361 as very useful. The DOR values of 16, 81, and 361 correspond to sensitivity-specificity pairs of 80%-80%, 90%-90%, and 95%-95%, respectively.

We used SAS for Windows, version 9.1.3 (SAS Institute) for all analyses and the PROC MIXED procedure in SAS to fit all models.

3.3 Results

Figure 3.1 summarizes how often the 7 QUADAS items were fulfilled in the 487 studies. Nonreporting of items was common, particularly for blinding of the index test (49%) and the reference test (72%), adequate time interval between the index and reference standard (42%), and whether patients were consecutively included (34%).

Nine of the 30 reviews included studies of the case-control or 2-gate type. Whether all patients had received the reference standard and whether the reference standard was the same for each patient were well reported (99% of the studies). In 3 reviews, the primary studies used different reference standards to verify index test results.

Applying the evidence-based definition of quality (items 5, 6, 10, and 11 of the QUADAS checklist) identified 72 (15%) of the 487 primary studies as high quality. With this definition, 12 of the 30 systematic reviews had no high-quality studies, and 9 reviews included at least 3 high-quality studies.

Applying the common-practice definition identified 70 high-quality studies (14%). With this definition, 9 systematic reviews contained no high-quality studies, and 11 reviews had at least 3 high-quality studies. Use of both definitions yielded only 3 reviews that contained 3 high-quality studies.

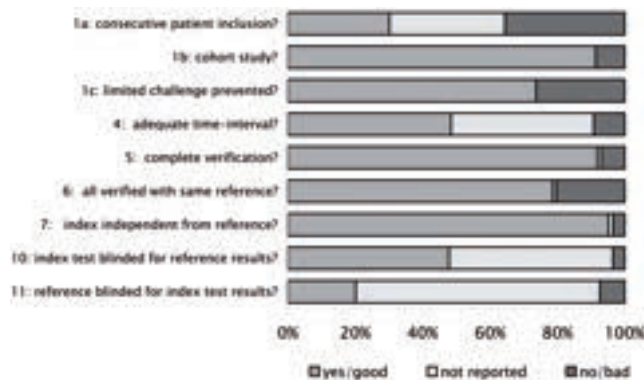


Figure 3.1. Overall results of quality assessment of the various QUADAS items in the 487 primary studies. Items 1a, 1b, and 1c refer to the different components of patient spectrum as we extracted them.

Figure 3.2. Point estimates of the DOR and confidence intervals of all analyses.
The abscissa represents the DOR, and the ordinate lists each meta-analysis by the first author, with the number of included studies in parentheses. Dotted lines reflect a DOR of 16 (i.e., a test with 80% sensitivity and 80% specificity), a DOR of 81 (90% sensitivity and 90% specificity) and DOR of 391 (95% sensitivity and 95% specificity). Analyses are indicated as follows: not incorporating quality (-), evidence-based restricted (■), common-practice restricted (□), evidence-based multivariable (▲), common-practice multivariable (Δ), and selective adjustment (x)

Table 3.3. Comparison of DORs and 95% confidence interval widths of different quality-incorporating strategies.^a

Alternative strategy (no. of analyses)	$H_0^b: \text{DOR}_{\text{overall}} = \text{DOR}_{\text{alternative}}$ Alternative DOR was:				$H_0: \text{CI}_{\text{overall}} = \text{CI}_{\text{alternative}}$ Alternative 95% CI was:			
	Higher	Lower	Equal	<i>P</i>	Broader	Smaller	Equal	<i>P</i>
Evidence-based restricted (9)	3	5	1	0.64	8	0	1	0.078
Common-practice restricted (11)	4	7	0	0.52	11	0	0	0.001
Evidence-based multivariable (21)	9	11	1	0.31	20	0	1	<0.001
Common-practice multivariable (23)	10	13	0	0.68	23	0	0	<0.001
Selective adjustment (30)	5	7	18	0.85	12	0	18	0.001

^a All strategies were compared with the overall meta-analysis, in which all studies within the original meta-analysis were included irrespective of their quality characteristics.

^b H_0 , null hypothesis; CI, confidence interval.

3.3.1 Comparing the pooled estimates of the various strategies

The summary DORs and the corresponding 95% confidence intervals were obtained for all 30 systematic reviews with the reference and 3 quality-adjustment strategies (Figure 3.2).

The evidence-based restrict strategy, which analyzed only high-quality studies according to the evidence-based definition, could be applied in 9 reviews containing 3 high-quality studies. In 3 cases, the DOR for the high-quality studies was higher than the DOR obtained by ignoring quality and including all studies, whereas the opposite occurred in 5 cases ($P = 0.64$). In 1 review, the DOR did not change, because all studies were high-quality studies according to the evidence-based definition. We found only 2 or fewer high-quality studies in the other reviews, and we did not calculate a summary estimate based on these small numbers.

The restrict strategy with the common-practice definition could be used in 11 reviews. This restrict strategy produced a higher DOR in 4 meta-analyses and a lower estimate in 7 others. The mean odds ratio was not significantly higher or lower when quality was not incorporated, compared with the different restrictive strategies (Table 3.3).

When we included all the items of the evidence-based quality definition as covariates in the multivariable model, model building failed in 9 reviews. In these reviews, at least 1 of the quality criteria was not fulfilled by any of the included studies. In 9 of the other 21 reviews, the adjust-all strategy resulted in a DOR estimate that was higher than when quality was not incorporated; 11 times the estimate was lower.

In 1 review, all of the original studies could be regarded as of high quality, so there was no change in the summary DOR.

With the common-practice definition, we were able to make a multivariable adjust-all model in 23 reviews. The estimated DOR was higher in 10 reviews and lower in 13. The differences between analyzing studies irrespective of their quality and analyses with the 2 multivariable strategies were not significant (Table 3.3).

The selective-adjustment strategy included only items that were significantly associated with accuracy in a univariable analysis ($P < 0.2$). In 18 reviews, none of the QUADAS items was significantly associated with accuracy, and the use of all original studies in a meta-analysis yielded the same summary DOR as when quality was disregarded. In 5 reviews, only one single QUADAS item had a significant effect, and in a further 5, 1, and 1 meta-analyses respectively 2, 3, and 4 items were significant. The selective-adjustment strategy led to a higher estimate in 5 cases and to a lower estimate in 7 cases, compared with the meta-analysis in which quality was not incorporated.

Figure 3.3 shows the relative DORs (compared with not including quality in the analysis) for the various quality-adjustment strategies. The symmetrical distribution around unity illustrates that there is no systematic trend in underestimating or overestimating the DOR of a test. However, in 5 cases, the alternative strategy resulted in a DOR >5 times higher than when quality was disregarded; in 3 cases the relative DOR was <0.2.

Figure 3.3. Relative DOR for each meta-analysis.

DORs of different quality-adjusting strategies are compared with the DOR for the ignore-quality strategy. A relative DOR >1.0 means that the DOR of the quality-adjusted meta-analysis was higher than when quality was not taken into account. A relative DOR <1.0 means that the DOR was greater when no adjustment for quality was made. The thin line represents a relative DOR of 1.0, i.e., no difference between the adjusted and nonadjusted analyses. Indicated are the evidence-based restricted strategy (■), the common-practice restricted strategy (▲), the evidence-based multivariable strategy (▼), the common-practice multivariable strategy (·), and the selective-adjustment strategy (●).

None of the quality-adjustment strategies produced systematically narrower confidence intervals for the summary DOR than analyzing studies irrespective of their quality (Table 3.3). The confidence intervals were significantly wider with the restrict and adjust-all strategies ($P < 0.01$) but did not significantly differ with the selective-adjustment method ($P = 0.08$).

Because differences between strategies can be due to both differences in quality definitions and differences in statistical methods, we compared the results between statistical methods within 1 definition. We also compared the results with 2 quality definitions within 1 strategy. We observed no systematic differences between the 2 approaches, either for the summary estimates or for their 95% confidence intervals.

The judgment about the usefulness of a test based on the magnitude of the summary DOR was not affected in 12 of the 30 reviews with any of the quality-adjustment strategies (Figure 3.2). In 18 reviews, the quality-adjusted DOR obtained with 1 or more of the quality-adjustment strategies ended in a different category than the DOR obtained with all studies included. The DOR was higher in 14 cases and lower in 17 others (Figure 3.2).

3.4 Discussion

In this re-analysis of 30 previously published systematic reviews, we found no evidence for our hypothesis that adjustment for differences in methodological quality in meta-analysis leads to less optimistic summary diagnostic accuracy estimates with less variability in results among better-quality studies. We saw no such overall effects for strategies that relied on restriction to high-quality subsets, on multivariable adjustment for a set of quality items, or on selective multivariable adjustment for significant quality items.

A main problem that authors of systematic reviews encounter is poor reporting of study characteristics, and our study was no exception⁵¹. We scored any study feature that was not reported as deficient. Dichotomizing QUADAS items into a simple “yes” or “no” can lead to loss of information, especially when many study characteristics are unreported. Some QUADAS items, such as the use of an adequate reference standard and the generalizability of the patient spectrum, could not be assessed at all in our data set. Both of these items can have a large effect on the performance of a test under study, and a proper incorporation of these characteristics could have resulted in a larger effect of the quality-adjustment strategies.

Because our analysis unit was the single meta-analysis, our sample size was only 30. Therefore, the power for detecting significant trends between strategies was limited, despite the inclusion of 487 individual studies. The 30 systematic reviews

covered a wide range of clinical topics and diagnostic tests, with a wide variability in the magnitude of the DOR. Our primary outcome variable was the DOR, which is a single accuracy indicator that incorporates both the sensitivity and specificity of a test. Such a single indicator is convenient in the analysis, but it also means that any given summary DOR can be produced by innumerable sensitivity-specificity combinations. In practice, the value of 1 accuracy measure, say sensitivity, may be more critical than another if the implications of false-positive and false-negative test results differ in severity.

In our analysis, we refrained from calculating summary quality scores for studies and labelling any study that exceeded a certain threshold score as high quality. Such summary quality scores have been extensively studied—and criticized—in systematic reviews of intervention studies. Different shortcomings in study design may cause different forms of bias, making it almost impossible to determine the weight that should be given to each quality item in calculating such quality scores^{52,53}. We also did not include a sequential analysis of the studies based on their quality ranking, which would have led to a quality-adjusted cumulative meta-analysis⁵⁴. This strategy also requires a hierarchical approach to study quality in that it assumes that some criteria are more important than others and that studies fulfilling more criteria are of higher quality.

Several previous studies have linked design features of diagnostic accuracy studies to changes in accuracy estimates. One systematic review documented the theoretical and empirical evidence for several sources of bias^{4,5}. Two publications, which examined these effects in a collection of systematic reviews, both reported significant effects for a number of features across meta-analyses^{1,2}. We can only speculate why we failed to find any systematic differences from incorporating these study features in the meta-analysis process. These earlier studies analyzed the impact of deficiencies in quality in a large number of diagnostic accuracy studies across a variety of systematic reviews, whereas our study assessed the impact of these quality items on estimates of diagnostic accuracy within systematic reviews. Furthermore, the number of studies with methodological deficiencies was small in a number of the systematic reviews included in our analysis, whereas other reviews contained only studies with deficiencies. Many of these studies with a deficient study design had a small sample size⁵⁵. Because the weight of an individual study depends on sample size, these studies had only a minor impact on the summary estimate of diagnostic accuracy. Furthermore, if 2 or more quality items influence accuracy but in opposing directions, the overall estimate obtained irrespective of quality may be similar to the estimate based on high-quality studies only. It is also possible that incomplete reporting has led to misclassification of design features in our project, which may have jeopardized our attempts to find differences in accuracy.

There are other potential explanations for our failed attempts at quality adjustment. The effects of several study-design features may not always be in the same predictable direction. Whether partial verification, for example, will lead to accuracy

estimates that are unchanged, lower, or higher, depends on the pattern of verification and the reference standards being used. The ratio of patients with unverified positive index test results and patients of unverified negative test results matters, in particular when being verified or not is related to the presence or absence of the target condition.

Similar remarks have been made in the field of intervention studies, where more meta-epidemiologic studies like ours have been performed^{56,57}. The aim in meta-epidemiologic studies is to evaluate the importance of 1 or more design features across a substantial number of systematic reviews. These studies have shown that meta-epidemiologic studies require substantial numbers of systematic reviews with sufficient differences in methodological quality among the included studies. Furthermore, if the effects of design features vary in direction among reviews or even among studies within a single review, meta-epidemiologic studies may produce summary estimates that suggest no effect at all^{58,59,60}. Although we have found no systematic trend in results among strategies, reviews in which adjusting for quality has led to substantially different results clearly exist. Because we do not know the true magnitude of accuracy, it is impossible to tell whether the adjusted estimates were closer to the truth.

Not only did we fail to find support for our hypothesis that adjusting for quality will result in less optimistic estimates of test accuracy, we also found no evidence for the hypothesis that adjusting for quality leads to less heterogeneity in results and therefore to smaller confidence intervals. On the contrary, the alternative analyses generally produced broader confidence limits. The main reason for this result is that the alternative strategies were based on fewer studies.

Our study did not produce evidence for the superiority of one type of adjustment over another. Low-quality studies can produce accuracy statistics that do not differ from those obtained in high-quality studies. Although methodological quality may influence the results of meta-analyses, a direct association with results is not necessarily present.

In any review, poor quality will affect the trustworthiness of the conclusions of that review. Our results indicate that the strategy used to correct for quality may affect the estimated accuracy, but not in a predictable way. Our results also indicate that measuring and incorporating quality in a diagnostic review is not a simple task of routinely scoring a few standard quality items and then adjusting for these variables in a multivariable model.

There may be good reasons to identify some quality criteria as crucial for the credibility and applicability of any systematic review. An example could be the selection of the reference standard—QUADAS item 3. These criteria may then be used as inclusion criteria for the review, and authors of systematic reviews might want to report how many studies had to be excluded based on that criterion.

Quality-assessment results of the studies included in a review remains a necessity because it notifies readers about the overall quality of the studies included in the review and may point out differences in design that can help to explain some of the heterogeneity in results. The QUADAS instrument can be used for that purpose. We propose to score “not reported” as a separate category where applicable, and we hope that a more widespread implementation of the STARD statement will lead to better reporting in future reports of diagnostic accuracy studies^{61,62}.

We feel it necessary that quality-assessment results in a systematic review be summarized in a table or a figure. A table can list the extent to which each of the studies fulfilled the quality criteria. A figure, such as the stacked bar chart in Figure 3.1, can then display the studies for which each of the respective criteria was fulfilled so that the reader can obtain an overview of the quality of the studies included in the review. Plotting results for all of the included studies in ROC space and coding individual studies by colour or with symbols can help readers recognize the characteristics of individual studies.

In our view, whether quality is also to be incorporated in a meta-analysis depends on several factors. In the first place, analyzing quality is not even an option if the number of included studies is too low. If the results are very heterogeneous, quality differences can be used to search for an explanation for the heterogeneity, and such a search can be accommodated by stratification or, if appropriate, regression analysis. Caution is needed because it is not unusual for the potential explanations for observed differences to outnumber the studies in a systematic review. It is important to recognize the major limitations of meta-epidemiologic approaches in meta-analysis.

Quality is a multidimensional concept, and the importance of individual quality items will vary from one research project to another. The goal of adjusting for quality differences in meta-analysis will remain attractive but elusive until we have large-scale systematic reviews and fully informative reporting in individual studies.

Acknowledgments

J.J.D. is supported in part by a Senior Scientist in Evidence Synthesis Award from the UK Department of Health.

References

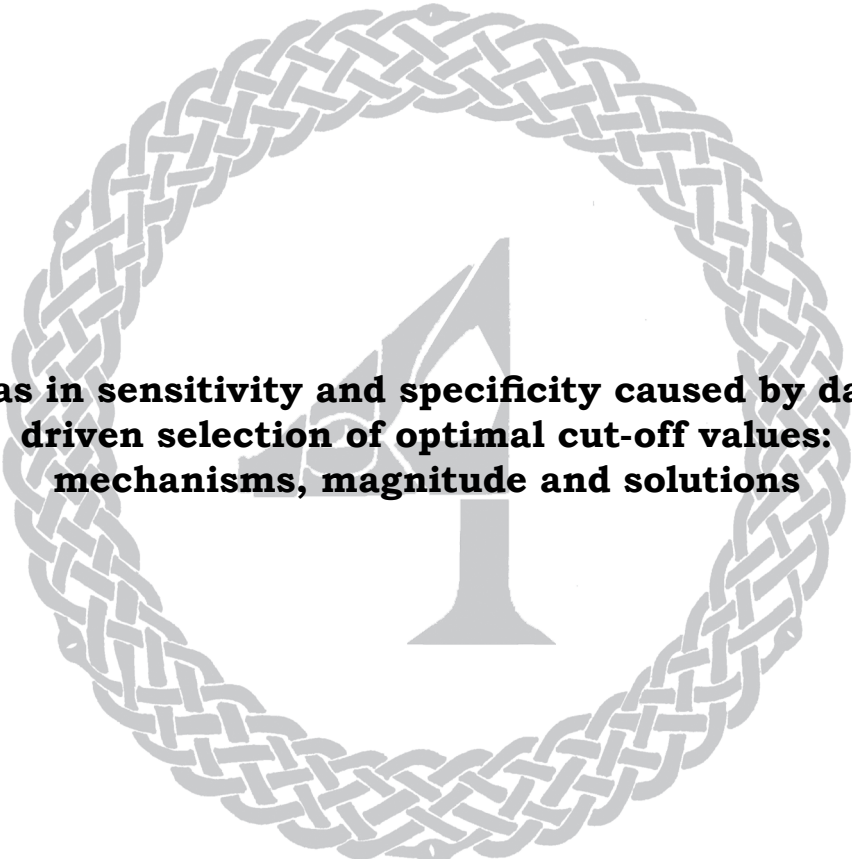
1. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999; 282(11):1061-6.
2. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, Van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ*. 2006; 174(4):469-76.
3. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? *BMC Med Res Methodol* 2005; 5(1):20.
4. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med*. 2004; 140(3):189-202.
5. Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess*. 2004; 8(25):1-234.
6. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Technol Assess* 2005; 9(12):1-128.
7. Lijmer JG, Bossuyt PM, Heisterkamp S. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med*. 2002; 21(11):1525-37.
8. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ*. 2002 Mar 16; 324(7338):669-71.
9. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003; 3:25.
10. Whiting PF, Westwood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol*. 2006; 6:9.
11. Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol*. 2005;58(1):1-12.
12. De Vet HC, van der WT, Muris JW, Heyrman J, Buntinx F, Knottnerus JA. Systematic reviews of diagnostic research. Considerations about assessment and incorporation of methodological quality. *Eur J Epidemiol* 2001;17(4):301-6.
13. Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol* 2002; 2:9.
14. Khan KS. Systematic reviews of diagnostic tests: a guide to methods and application. *Best Pract Res Clin Obstet Gynaecol* 2005; 19(1):37-46.
15. Balk EM, Ioannidis JP, Salem D, Chew PW, Lau J. Accuracy of biomarkers to diagnose acute cardiac ischemia in the emergency department: a meta-analysis. *Ann Emerg Med* 2001; 37(5):478-94.
16. Berger MY, van der Velden JJ, Lijmer JG, de Kort H, Prins A, Bohnen AM. Abdominal symptoms: do they predict gallstones? A systematic review. *Scand J Gastroenterol* 2000; 35(1):70-6.
17. Deville WL, van der Windt DA, Dzaferagic A, Bezemer PD, Bouter LM. The test of Lasegue: systematic review of the accuracy in diagnosing herniated discs. *Spine* 2000; 25(9):1140-7.
18. Fiellin DA, Reid MC, O'Connor PG. Screening for alcohol problems in primary care: a systematic review. *Arch Intern Med* 2000; 160(13):1977-89.

19. Gould MK, Maclean CC, Kuschner WG, Rydzak CE, Owens DK. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: a meta-analysis. *JAMA* 2001; 285(7):914-24.
20. Hobby JL, Tom BD, Bearcroft PW, Dixon AK. Magnetic resonance imaging of the wrist: diagnostic performance statistics. *Clin Radiol* 2001; 56(1):50-7.
21. Hoffman RM, Clanon DL, Littenberg B, Frank JJ, Peirce JC. Using the free-to-total prostate-specific antigen ratio to detect prostate cancer in men with nonspecific elevations of prostate-specific antigen levels. *J Gen Intern Med* 2000; 15(10):739-48.
22. Hoogendam A, Buntinx F, de Vet HC. The diagnostic value of digital rectal examination in primary care screening for prostate cancer: a meta-analysis. *Fam Pract* 1999; 16(6):621-6.
23. Huicho L, Campos-Sanchez M, Alamo C. Metaanalysis of urine screening tests for determining the risk of urinary tract infection in children. *Pediatr Infect Dis J* 2002; 21(1):1-11.
24. Hurley JC. Concordance of endotoxemia with gram-negative bacteremia. A meta-analysis using receiver operating characteristic curves. *Arch Pathol Lab Med* 2000; 124(8):1157-64.
25. Kelly S, Harris KM, Berry E, Hutton J, Roderick P, Cullingworth J et al. A systematic review of the staging performance of endoscopic ultrasound in gastro-oesophageal carcinoma. *Gut* 2001; 49(4):534-9.
26. Kim C, Kwok YS, Heagerty P, Redberg R. Pharmacologic stress testing for coronary disease diagnosis: A meta-analysis. *Am Heart J* 2001; 142(6):934-44.
27. Koelemay MJ, Lijmer JG, Stoker J, Legemate DA, Bossuyt PM. Magnetic resonance angiography for the evaluation of lower extremity arterial disease: a meta-analysis. *JAMA* 2001; 285(10):1338-45.
28. Kwok Y, Kim C, Grady D, Segal M, Redberg R. Meta-analysis of exercise testing to detect coronary artery disease in women. *Am J Cardiol* 1999; 83(5):660-6.
29. Lau J, Ioannidis JP, Balk EM, Milch C, Terrin N, Chew PW et al. Diagnosing acute cardiac ischemia in the emergency department: a systematic review of the accuracy and clinical effect of current technologies. *Ann Emerg Med* 2001; 37(5):453-60.
30. Lederle FA, Simel DL. The rational clinical examination. Does this patient have abdominal aortic aneurysm? *JAMA* 1999; 281(1):77-82.
31. Li J. Capnography alone is imperfect for endotracheal tube placement confirmation during emergency intubation. *J Emerg Med* 2001; 20(3):223-9.
32. Mitchell MF, Cantor SB, Brookner C, Utzinger U, Schottenfeld D, Richards-Kortum R. Screening for squamous intraepithelial lesions with fluorescence spectroscopy. *Obstet Gynecol* 1999; 94(5 Pt 2):889-896.
33. Mol BW, Lijmer JG, van der MJ, Pajkrt E, Bilardo CM, Bossuyt PM. Effect of study design on the association between nuchal translucency measurement and Down syndrome. *Obstet Gynecol* 1999; 94(5 Pt 2):864-9.
34. Nelemans PJ, Leiner T, de Vet HC, van Engelshoven JM. Peripheral arterial disease: meta-analysis of the diagnostic performance of MR angiography. *Radiology* 2000; 217(1):105-14.
35. Safriel Y, Zinn H. CT pulmonary angiography in the detection of pulmonary emboli: a meta-analysis of sensitivities and specificities. *Clin Imaging* 2002; 26(2):101-5.
36. Sloan NL, Winikoff B, Haberland N, Coggins C, Elias C. Screening and syndromic approaches to identify gonorrhoea and chlamydial infection among women. *Stud Fam Plann* 2000; 31(1):55-68.
37. Smith-Bindman R, Hosmer W, Feldstein VA, Deeks JJ, Goldberg JD. Second-trimester ultrasound to detect fetuses with Down syndrome: a meta-analysis. *JAMA* 2001; 285(8):1044-55.

38. Sonnad SS, Langlotz CP, Schwartz JS. Accuracy of MR imaging for staging prostate cancer: a meta-analysis to examine the effect of technologic change. *Acad Radiol* 2001; 8(2):149–57.
39. Vasquez TE, Rimkus DS, Hass MG, Larosa DI. Efficacy of morphine sulfate-augmented hepatobiliary imaging in acute cholecystitis. *J Nucl Med Technol* 2000; 28(3):153–5.
40. Visser K, Hunink MG. Peripheral arterial disease: gadolinium-enhanced MR angiography versus color-guided duplex US—a meta-analysis. *Radiology* 2000; 216(1):67–77.
41. Westwood ME, Kelly S, Berry E, Bamford JM, Gough MJ, Airey CM et al. Use of magnetic resonance angiography to select candidates with recently symptomatic carotid stenosis for surgery: systematic review. *BMJ* 2002; 324(7331):198.
42. Wiese W, Patel SR, Patel SC, Ohl CA, Estrada CA. A meta-analysis of the Papanicolaou smear and wet mount for the diagnosis of vaginal trichomoniasis. *Am J Med* 2000; 108(4):301–8.
43. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem*. 2005; 51(8):1335–41
44. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993; 13(4):313–21.
45. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993; 12(14):1293–1316.
46. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995; 48(1):119–130.
47. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* 2002; 21(4):589–624.
48. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med*. 2004; 23(9):1351–75.
49. Steyerberg EW, Eijkemans MJ, Van Houwelingen JC, Lee KL, Habbema JD. Prognostic models based on literature and individual patient data in logistic regression analysis. *Stat Med*. 2000; 19(2):141–60.
50. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med*. 2000;19(8):1059–79.
51. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, Bouter LM, de Vet HC. Quality of reporting of diagnostic accuracy studies. *Radiology*. 2005; 235(2):347–53.
52. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999; 282(11):1054–60.
53. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol*. 2005; 5:19.
54. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol* 1992; 45(3):255–65.
55. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med*. 2001; 135(11):982–9.
56. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG; International Stroke Trial Collaborative Group; European Carotid Surgery Trial Collaborative Group. Evaluating non-randomised intervention studies. *Health Technol Assess*. 2003; 7(27):1–173.
57. Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med*. 2002; 21(11):1513–24.

58. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, Lau J. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA*. 2002; 287(22):2973–82.
59. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998; 352(9128):609–13.
60. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; 273(5):408–12.
61. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ*. 2003; 326(7379):41–4.
62. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, Bouter LM, de Vet HC. Reproducibility of the STARD checklist: an instrument to assess the quality of reporting of diagnostic accuracy studies. *BMC Med Res Methodol*. 2006; 6:12.





**Bias in sensitivity and specificity caused by data
driven selection of optimal cut-off values:
mechanisms, magnitude and solutions**

**Mariska M.G. Leeflang, Karel G.M. Moons,
Johannes B. Reitsma, Aeilko H. Zwinderman**

Clin Chem. 2008; 54(4):729-37

Abstract

Background: Optimal cut-off values for continuous test results are often derived in a data-driven way. This may however lead to overoptimistic measures of diagnostic accuracy.

Aim of study: To determine the magnitude of bias in sensitivity and specificity associated with data-driven selection of cut-off values and to examine potential solutions to reduce this bias.

Methods: Simulation study using different sample sizes, distributions and prevalences. We compared data-driven estimates of accuracy based on the Youden index with the true values, and calculated the median bias. Three alternative approaches (assuming specific distribution, leave-one-out, smoothed ROC) were examined for their ability to reduce this bias.

Results: The magnitude of bias caused by data-driven optimization of cut-off values was inversely related to sample size. If the true value of sensitivity and specificity are 84%, estimates in studies with a total sample size of 40 will be around 90%. If sample size increases to 200, estimates will be 86%. The distribution of the test results had little impact on the amount of bias if sample size was held constant. More robust methods of optimizing cut-off values were less prone to bias, but the performance deteriorated if the underlying assumptions were not met.

Discussion: Data-driven selection of the optimal cut-off value can lead to overoptimistic estimates of sensitivity and specificity, especially in small studies. Alternative methods can reduce this bias, but finding robust estimates of cut-off values and accuracy requires considerable sample sizes.

4.1 Introduction

Diagnostic accuracy is the amount of agreement between the results of an index test (the test under evaluation) and the reference standard (the best available method to determine the presence or absence of the disease of interest). Commonly used accuracy measures are sensitivity (the proportion of those with the target condition who have a positive index test result) and specificity (the proportion of those without the target condition who have a negative index test result). In case of a continuous or ordinal test, the ROC curve is an informative way to present the sensitivity versus 1–specificity for each possible cut-off value of the index test^{1,2}. In situations where higher test results are more indicative of the presence of disease, lowering the cut-off value will increase sensitivity, while specificity decreases. For clinical purposes in order to link actions to test results, one threshold or cut-off value is used. The optimal choice of this cut-off value is ultimately determined by the consequences associated with false positive and false negative test results³.

In early phases of test development, when the exact role of the index test is not fully defined and thus the consequences of incorrect test results are not yet determined, a criterion that equally weighs both sensitivity and specificity is often preferred to choose the optimal cut-off value. Such a criterion is the Youden index, which is defined by sensitivity + specificity – 1^{4,5}. The optimal cut-off value that maximizes the Youden index is often determined in a “data-driven” way. This means that the sensitivities and specificities across all possible cut-off values within the range of test results are calculated from the data at hand, and the cut-off value that leads to the highest Youden index is then selected.

This data-driven selection of optimal cut-off values is prone to bias, meaning that it systematically leads to overestimation of sensitivity and specificity of the test under study. Because chance variation plays a larger role in smaller studies, it means that the observed ROC curve from a single small study will deviate more from the true underlying ROC curve than the observed ROC curve from a large study (see Figure 4.1). These fluctuations occur in both directions leading to both underestimation and overestimation in relation to the true sensitivity and specificity. In small studies, an increase in sensitivity by taking a lower threshold will not directly lead to a decrease in specificity. Because the data-driven approach specifically selects the cut-off value with the highest sum of sensitivity and specificity (i.e. closest to the top left corner of the ROC plot), it is generally a point above the true underlying ROC curve. Data-driven selection of cut-off values for continuous test results in studies with low sample size may therefore lead to overoptimistic estimates of sensitivity and specificity. Because small sample sizes (<200) are common in diagnostic studies⁶, overestimation of diagnostic accuracy by data-driven selection of cut-off values can be a serious and prevalent problem.

This potential for bias associated with data-driven selection of the optimal Youden index has been recognized before, both in diagnostic and prognostic studies⁷⁻¹³.

Figure 4.1. ROC curves from three single studies in which the test results have been generated from the same underlying distribution, but with varying sample size. Disease prevalence was 50% in all three studies. The dashed line is based on a single study with a total sample size of 40 patients; the dotted line on a study with 1000 patients; the solid line is the true ROC curve belonging to a study with an infinite number of patients. The data-driven maximum Youden indices for the two empirical datasets are pointed by arrows: the upper arrow points at the optimal cut-off value in the population with 40 patients and the lower arrow points at the optimal cut-off value for sample size 1000. The true optimal sensitivity and specificity are both 84%.

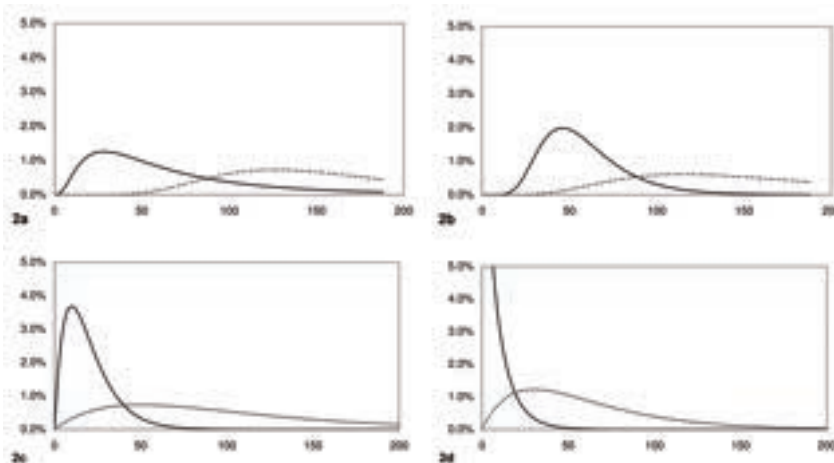


Figure 4.2. Alternative distributions. Alternative distributions of generating non-normal test results in our simulations: two Lognormal distributions (4.2a and 4.2b) and two Gamma distributions (4.2c and 4.2d). The solid lines reflect the distribution within the patients without the disease and the dotted lines in the patients with the disease. On the X-axis the test result and on the Y-axis the percentage of patients with that test result.

These publications have been rather technical, without offering clear guidance or solutions for practice. We therefore performed a series of simulations to document the magnitude of overestimation of sensitivity and specificity under a range of conditions and examined the possible role of alternative ways of estimating the sensitivity and specificity, using the Youden index. Based on these simulations we will be able to inform readers when to be aware of this bias and advice researchers how to reduce the potential for this bias in future studies.

4.2 Methods

4.2.1 Simulation of data sets

Continuous index test results for individuals with and without the disease were simulated based on a specific distribution, sample size and disease prevalence. The true values of optimal Youden index and cut-off and the corresponding true maximum sensitivities and specificities, were calculated from the true underlying distribution of test results among individuals with the disease and those without the disease.

To examine the impact of sample size, disease prevalence, amount of spread in index test results, and their underlying distribution on the amount of bias, we varied these parameters across scenarios. Sample sizes varied from 20 to 1000 patients; prevalence from 5% to 95%; standard deviations from 5 to 20; and test results were generated from an underlying Normal distribution and two non-symmetrical distribution Lognormal and Gamma distribution (see Figure 4.2).

All analyses were carried out using SAS for Windows, version 9.1.3 (SAS Institute).

4.2.2 Data driven estimation of sensitivity and specificity

Data driven estimates of diagnostic accuracy associated with the optimal cut-off value were determined in each simulated data set and compared with their true values. Each simulation scenario was replicated 2000 times to determine the median magnitude of bias (difference between each data driven estimate of both sensitivity and specificity and their true values) and the number of times (%) sensitivity and specificity were overestimated.

4.2.3 Potential solutions to reduce overestimation

Three alternative methods were examined whether they can reduce the magnitude of bias: (a) using sample characteristics and assuming a specific underlying distribution, (b) leave-one-out cross-validation, (c) robust fitting of ROC-curves. These methods were applied to two scenarios with a true underlying distribution of the index test results that was a Normal distribution, two scenarios with a true underlying Lognormal distribution and two scenarios with a true underlying Gamma distribution (see Figure 4.2). Within each simulated data set we compared the data

driven estimate with the estimates of the potential solutions to examine the effectiveness of the solutions in reducing the bias.

Deriving optimal cut-off point from assumed distributions

Sample characteristics describing the central tendency and shape of distribution of test results can be used to estimate the optimal cut-off value. By assuming a specific underlying distribution (e.g. a Normal distribution) for the test results in the patients with the disease, these sample characteristics (descriptives like mean and SD) can be used to calculate the cumulative proportion of diseased patients who will have an index test result equal to or above that cut-off value, e.g. an estimate of true sensitivity. Similarly, using the observed mean and SD of the non-diseased patients, the proportion of patients without the disease and with an index test result below each possible cut-off value can be calculated. This equals the specificity of that test. The Gamma distribution is characterized by a shape and a scale parameter, that, just like the mean and SD, describe the variation in test results in individuals with and without the disease within a sample. The lower the shape parameter, the more skewed the distribution is. The lower the scale parameter, the less spread the results are (just like a smaller standard deviation in Normal distributions). We estimated the shape and scale parameters of a Gamma distribution, based on the sampled data, using the Univariate Procedure. The cumulative Gamma distribution was then used to calculate sensitivity and specificity.

Leave-one-out cross-validation

In the leave-one-out cross validation a single subject is removed from the study population and used in the validation process. In the remaining (n-1) subjects the cut-off value is determined in a data-driven way, as described above. Thereafter, the resulting cut-off is applied to the single subject that did not take part in this process. This subject is then classified as either true positive, false positive, false negative, true negative depending on whether the subject is classified as having or not having the disease and whether its test result is below or above the cut-off value. This process is repeated for all patients in the data set and the resulting 2-by-2 table based on all subjects is used to determine sensitivity and specificity corresponding to the cut-off value which was derived in the n-1 patients.

Robust ROC curve fitting

In the robust ROC fitting approach a smooth, non-parametric curve is fitted through the observed data points plotted in ROC through a smoothing procedure which is included in SAS software (LOESS Procedure). The point on the fitted curve with the highest Youden index was used to obtain estimates of sensitivity and specificity.

4.2.4 Empirical evidence from published diagnostic reviews

From a set of 28 published systematic reviews, used in a previously published meta-epidemiological project, we selected those reviews that reported on continuous test results and included both studies with and without a pre-specified cut-off value. We then compared the summary diagnostic odds ratio between those two groups to ex-

amine whether the diagnostic accuracy was higher (overestimated) in studies with data driven selection of cut-off values than in studies using pre-specified cut-off values. The diagnostic odds ratio is an overall measure of accuracy combining both sensitivity and specificity: $[\text{sens}/(1-\text{sens})] \backslash [(1-\text{spec})/\text{spec}]$. Further details about this set of systematic reviews and the applied statistical methods can be found in an earlier publication¹⁴.

4.3 Results

4.3.1 Simulation of data sets

In the basic scenario, index test results were generated from a Normal distribution with a mean value of 100 (SD=10) for persons without the disease and a mean value of 120 (SD=10) for persons with the disease, leading to a true maximum Youden index of 0.68, a true optimal cut-off value of 110 and true values of sensitivity and specificity of both 84%. These true values will only alter if the underlying distribution changes (like the difference in means between diseased and non-diseased or the spread of test results), but are not affected by changes in sample size or disease prevalence.

Figure 4.3. Effect of sample size on data-driven estimates of sensitivity.

The median sensitivity across 2000 simulations together with the 25th and 75th percentiles are shown. On the X-axis: the number of persons with the disease (disease prevalence was 50%). Data based on Normally distributed test results with mean=100 and SD=10 for non-diseased and mean=120 and SD=10 for diseased. The dotted line represents the true value of sensitivity. The results for specificity were similar.

4.3.2 Data driven overestimation of sensitivity and specificity

The effect of sample size

The amount of bias in the data-driven estimates was inversely related to sample size (Figure 4.3). At a total sample size of 40, the median sensitivity and specificity were both 90% (interquartile range 80 to 95%), while their true values were both 84%. Both measures were overestimated in 74% of the simulations. In sixty percent of the simulations, estimates of sensitivity and specificity exceeded 89%, while their true value was 84%. When the total sample size was 200, sensitivity was overestimated in 62% and specificity in 60% of all simulations, while their median values were approaching their true values (86% (interquartile range 82 to 89%) in stead of 84%).

The effect of disease prevalence

A prevalence of 50% is the most efficient prevalence to ensure that combined uncertainty in both sensitivity and specificity is smallest. This was also reflected in our results. Lowering the prevalence (conditional on the same total sample size) leads to fewer individuals with the disease, larger fluctuation in sensitivity by chance and therefore more room for overestimation of sensitivity. The opposite occurs for specificity. The median absolute bias at a prevalence of 10% was 5.9% for sensitivity and 3.6% for specificity. At a prevalence of 90%, the median absolute bias was 2.2% for sensitivity and 6.7% for specificity (results not shown).

Overlap in test results between populations with and without the disease

The spread and overlap in test results between populations with and without the disease determines the absolute size of sensitivity and specificity. A smaller standard deviation (less spread) while the difference in mean values between the populations remains the same, will lead to less overlap in test results between diseased and non-diseased. Thus, sensitivity and specificity will increase, leaving less room for overestimation (ceiling effect): sensitivity cannot exceed 100%. On the other hand, if we allow the standard deviations to change without changing sensitivity and specificity, then the amount of bias did not vary (results not shown).

The effect of underlying distributions

The underlying distribution of the simulated test results by comparing scenarios based on a Normal, Lognormal or Gamma distribution had little impact on the average amount of bias (see Figures 4.4 and 4.5). However, the amount of bias could vary substantially within a specific distribution based on the actual values of the parameters of that distribution. For example, one of the Lognormal distributions resulted in 60% of the simulations with an overestimation of sensitivity that was more than 5% points, while the other Lognormal distribution resulted in such an overestimation in 35% of the simulations.

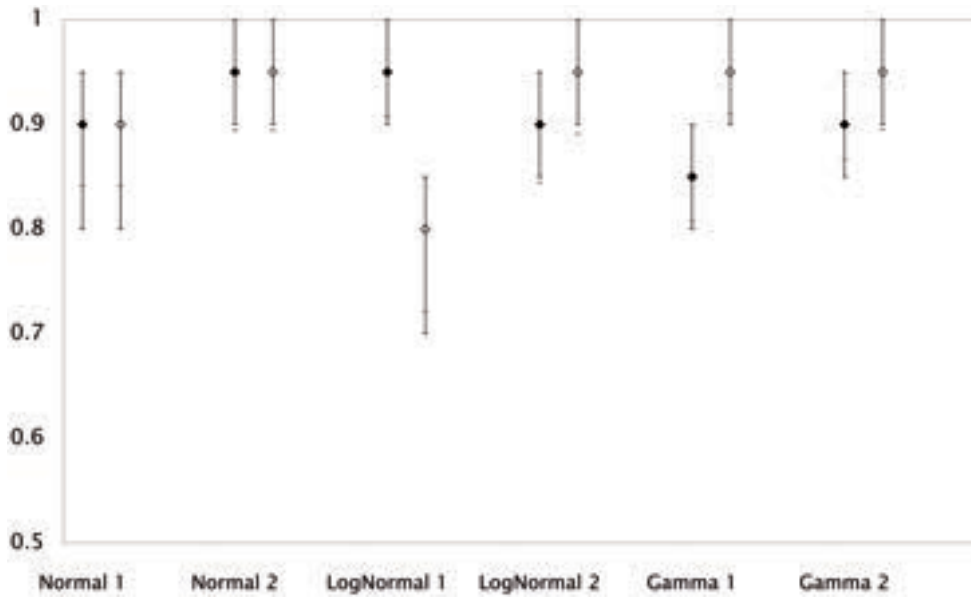


Figure 4.4. The effect of the underlying distribution on sensitivity and specificity.
 The closed diamonds are the median data driven values of the sensitivities and the open diamonds are the median values of the specificities. Also shown are the data-driven 25th and 75th percentiles and the true values (dashes). Prevalence was in all situations 50% and total sample size was 40. Normal distribution 1: mean(SD) diseased = 120(10) and mean(SD) non-diseased = 100(10). Normal distribution 2: mean(SD) diseased = 122.5(10) and mean(SD) non-diseased = 97.5(10). The Lognormal and Gamma distributions are shown in Figure 4.2.

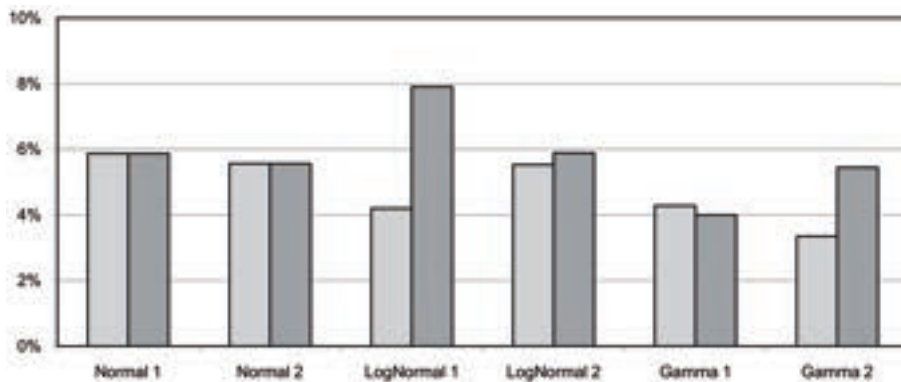


Figure 4.5. Effect of the underlying distribution on the absolute amount of bias in sensitivity (light) and specificity (dark).
 Prevalence was in all situations 50% and total sample size was 40. On the Y-axis the absolute bias in % points above the true value. Normal distribution 1: mean(SD) diseased = 120(10) and mean(SD) non-diseased = 100(10). Normal distribution 2: mean(SD) diseased = 122.5(10) and mean(SD) non-diseased = 97.5(10). The Lognormal and Gamma distributions are shown in Figure 4.2.

4.3.3 Potential solutions to reduce bias

Deriving optimal cut-off point from assumed distributions

Using the estimated mean and standard deviation from a data set and then calculating the true optimal cut-off value by assuming a Normal distribution, decreases the amount of bias when the underlying distribution was indeed Normal. In one of the scenarios with a true underlying Normal distribution of the index test results, median sensitivity and specificity following this strategy were both 85%, while their true value was 84%. This is a difference of only 1% point (see Figure 4.6).

When the underlying distribution is a Gamma or Lognormal one, this same procedure leads to a systematic underestimation of sensitivity and overestimation of specificity, which was sometimes worse than the uncorrected, data-driven results. In these situations, the median estimated sensitivity was 2-13% points lower than the true sensitivity (see Figure 4.6). The difference between the median estimated specificity and the underlying true specificity was 7 or 8% points.

The Gamma distribution is more flexible in approximating various distributions and led to less bias in all scenarios than the data-driven method. The median estimated sensitivity varied from 2% points below to 3% points above the true sensitivity. The median estimated specificity varied from 1% points to 4% points above the true specificity.

Because we sometimes observed that results for sensitivity and specificity were in the opposite direction (overestimation in one and underestimation in the other parameter), we summed the absolute value of the bias in sensitivity and specificity. When we assumed the underlying distributions to be Normal, the total absolute value of bias was 59% points (summed absolute bias of all sensitivities in all five studied scenarios was 28% points, summed bias of all specificities in all five sce-

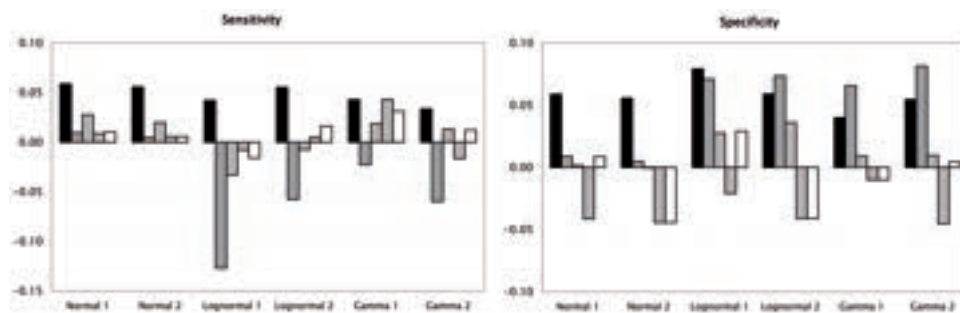


Figure 4.6. Figures 4.6a, bias in sensitivity, and 4.6b, bias in specificity.

Median amount of bias of the data-driven and alternative approaches in relation to the true value. On the Y-axis the absolute amount of bias in % points above or under the true value. Disease prevalence was 50% in all situations and total sample size was 40. Black bars, data-driven analysis; grey bars, assuming a Normal distribution; top-left-to-bottom-right striped bars, assuming a Gamma distribution; top-right-to-bottom-left striped bars, leave-one out validation; white bars, robust ROC fitting.

narios was 31% points). When we assumed underlying Gamma distributions, the total absolute value of bias was 20% points (sum of absolute bias in sensitivity was 12% points and in specificity 8% points).

Leave-one-out cross-validation

The leave-one-out cross validation resulted in less bias in sensitivity, the estimated values were 2% lower to 4% higher than their true values. Specificity was a marginally underestimated (1 to 5% lower than their true value) (see Figure 4.6). The sum of the absolute value of the bias in sensitivity was 9% points and in specificity 20% points (total bias of 29% points).

Robust ROC curve fitting

Robust fitting of ROC curves also resulted in less bias in both sensitivity and specificity. Difference between true and estimated sensitivity ranged from minus 2% points to 3% points and difference between true and estimated specificity ranged from minus 4% points to 3% points (see Figure 4.6). The sum of the absolute value of the bias in sensitivity was 9% points and in specificity 14% points (total bias of 23% points).

4.3.4 Empirical evidence from published diagnostic reviews

A total of seven systematic reviews evaluated a test producing continuous test results and five of these reviews included both studies with a pre-specified cut-off value and studies with a data driven cut-off value. The diagnostic odds ratio on average was 1.71 (95% confidence interval: 1.04 to 2.82; $P=0.03$) times higher in studies with a data-driven cut-off value compared to studies with a pre-specified cut-off value. Translating this results to sensitivity and specificity, it means that if a study with a pre-specified cut-off value would estimate sensitivity and specificity both at 84% (=diagnostic odds ratio of 28), a study using data-driven selection would find estimates of sensitivity and specificity of 87.4% , corresponding to a diagnostic odds ratio of 48 (=28 times 1.71).

4.4 Discussion

Our simulation study showed that data-driven selection of the optimal cut-off values for a continuous test by using the Youden index led to overestimated estimates of sensitivity and specificity. The amount of bias in sensitivity and specificity was predominantly dependent on total sample size. A typical value for the absolute amount of bias in studies with a sample size of 40 was 5% points occurring in both sensitivity and specificity.

The amount of bias becomes smaller by increasing sample size. Overestimation of more than 5% was present in 27% of the simulations if the total sample size was 200 compared to 60% of the studies with sample size of 40. The underlying distri-

butions had little or no effect on the amount of bias. This can be explained by the non-parametric way of data driven selection of the optimal cut-off value. The absolute magnitude of the true sensitivity and specificity did have an effect: the nearer the true values approached 100%, the less room there was for overestimation.

In this study, we have only reported the effect of optimizing cut-off values on sensitivity and specificity, although we also examined the effects on likelihood ratios and diagnostic odds ratios (results not reported). These effects were in line with the results on sensitivity and specificity, which is not surprising because they are direct functions of sensitivity and specificity. This potential for bias was confirmed in our empirical data, as the diagnostic odds ratio in studies with data-driven cut-off values was significantly higher than in studies with pre-specified values.

We applied three alternative and more robust methods for determining the sensitivity and specificity associated with the optimal cut-off value to examine whether these methods were less prone to bias. In general, these methods resulted in lower estimates of sensitivities and specificities, sometimes even producing too conservative estimates (see Figure 4. 6). As expected, the performance of the method which assumes that the underlying distribution was Normal deteriorated considerably if this assumption was not met. Because it is difficult to examine in a small sample whether it is reasonable to assume a Normal underlying distribution, we do not recommend this method in general. Assuming a Gamma distribution is a more flexible approach, as it can mimic various shapes of distribution and therefore this method performed consistently well across our simulations. The smooth ROC fitting can be viewed as a distribution-free method, meaning that it would perform consistently irrespective of the true underlying distribution. The leave-one-out approach is a traditional way of cross validating the results in regression analyses to reduce the impact of over fitting. In our situation, the leave-one-out approach produced indeed lower estimates than the data-driven method. However, sometimes the estimates from the leave-one-out approach became too conservative, especially for specificity. We do not have an explanation for this. Bootstrapping would have been a slightly different approach based on the same principle of cross-validation. Therefore, we expect similar results with this method as with the leave-one-out approach.

Another approach that will reduce the problem of overestimation is using a pre-specified cut-off value. However, in the early phase of test evaluation, there may be little indication about the likely value of the optimal cut-off value. Other more complex solutions to generate less biased results, but still use the actual data of a study have been described. These involve the reporting of a confidence interval around the 'true' cut-off value and a Bayesian method to smooth the steps in an ROC curve. Details can be found here^{5,15}.

Readers of diagnostic studies should be aware of the potential for bias when optimal cut-off values have been derived in a data-driven way, especially if the sample size was small. Defining a small study is rather arbitrary and depends on the amount

of bias you are willing to accept. Our results show that there is probability of 27% that sensitivity and specificity will be overestimated more than 5% points in a study with a sample size 200. A rule of thumb would be that a diagnostic study should have at least 100 individuals without the disease as well as 100 individuals with the disease before a cut-off value can be reliably estimated from the data. The problem is however, that most diagnostic studies will not have these numbers⁶. Another problem both clinicians and laboratory professionals may encounter, is that not only the amount of bias will increase if sample sizes get smaller, also the confidence interval around the estimate of the optimal cut-off value and of both sensitivity and specificity will increase. Even if bias is reduced by using more robust methods, uncertainty about the true optimal cut-off value and its corresponding diagnostic accuracy will remain.

In conclusion, researchers and readers of diagnostic studies should be aware of over optimistic measures of diagnostic accuracy when the results have been generated by a data-driven approach in a small study. Several methods exist that can reduce the amount of this bias, but it is important to stress that finding robust estimates of cut-off values and their associated measures of accuracy require studies of considerable sample size. In smaller studies, researchers may present a scatter graph showing the distribution of all test results in the non-diseased and the diseased individuals. In addition they can draw the empirical ROC curve and a robust (smoothed) ROC curve, but refrain from selecting the most outlying point closest to the top left corner (=maximum Youden).

References

1. Shapiro DE. The interpretation of diagnostic tests. *Stat Meth Med Res.* 1999; 8(11):113–34.
2. Obuchowski NA, Lieber ML, Wians FH Jr. ROC curves in Clinical Chemistry: uses, misuses, and possible solutions. *Clin Chem.* 2004; 50(7):1118–25.
3. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver–operating characteristic analysis for diagnostic tests. *Prev Vet Med.* 2000; 45(1–2):23–41.
4. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950; 3(1):32–5.
5. Fluss R, Faraggi D, Reiser B. Estimation of the Youden index and its associated cutoff point. *Biometrical J.* 2005; 47(4):458–72.
6. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ.* 2006; 332(7550):1127–9.
7. Linnet K, Brandt E. Assessing diagnostic tests once an optimal cutoff point has been selected. *Clin Chem.* 1986; 32(7):1341–46.
8. Le CT. A solution for the most basic optimization problem associated with an ROC curve. *Stat Methods Med Res.* 2006; 15(6):571–84.
9. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. *J Clin Epidemiol.* 2006; 59(8):798–801.
10. Jund J, Rabilloud M, Wallon M, Ecochard R. Methods to estimate the optimal threshold for normally or log–normally distributed biological tests. *Med Decis Making.* 2005; 25(4): 406–15.
11. Perkins NJ, Schisterman EF. The Youden index and the optimal cut–point correctment for measurement error. *Biometrical J.* 2005; 47(7):428–441.
12. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut–point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology.* 2005; 16(1):73–81.
13. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using “optimal” cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst.* 1994; 86(11):829–35.
14. Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deeks J, Bossuyt P. Impact of adjustment for quality on results of meta–analyses of diagnostic accuracy. *Clin Chem.* 2007; 53(2):164–72.
15. Gail MH, Green SB. A generalization of the one–sided two–sample Kolmogorov–Smirnov statistic for evaluating diagnostic tests. *Biometrics* 1976; 32(3):561–570.





**Diagnostic accuracy may vary with prevalence:
Implications for evidence-based diagnosis**

Mariska M.G. Leeflang, Patrick M.M. Bossuyt, Les Irwig

Accepted by J Clin Epidemiol

Abstract

Background: Sensitivity and specificity of diagnostic tests are often assumed to be independent of prevalence. Yet several studies and systematic reviews have reported results that indicate otherwise.

Methods: We identify and explore mechanisms that may be responsible for sensitivity and specificity varying with prevalence and illustrate them with examples from the literature.

Results: Clinical and artefactual variability may be responsible for changes in prevalence and accompanying changes in sensitivity and specificity. Clinical variability refers to differences in the clinical situation that may cause sensitivity and specificity to vary with prevalence. For example, a patient population with a higher disease prevalence may include more severely diseased patients, in which the test performs better. Artefactual variability refers to effects on prevalence and accuracy associated with study design, for example the verification of index test results by a reference standard. Changes in prevalence influence the extent of overestimation due to imperfect reference standard classification.

Conclusions: Sensitivity and specificity may vary in different clinical populations, and prevalence is a marker for such differences. Clinicians are advised to base their decisions on studies that most closely match their own clinical situation, using prevalence to guide the detection of differences in study population or study design.

5.1 Introduction

Diagnostic test accuracy refers to the ability of a test to discriminate between those who have and those who do not have the target condition. Accuracy is assessed by comparing the results of the index test, the test under evaluation, with the results of the reference standard, which aims to classify patients as having or not having the target condition. Test accuracy is most often expressed as the test's sensitivity (the proportion of those with the target condition who have a positive index test result) and specificity (the proportion of those without the target condition who have a negative index test result).

A test's sensitivity and specificity are commonly believed not to vary with disease prevalence. Yet a number of studies have shown that differences in diagnostic accuracy often accompany differences in prevalence between study groups (see Table 5.1 for examples). For example, Flicker and colleagues used a consensus diagnosis as the reference standard in assessing the diagnostic accuracy of different checklists for dementia. They found a lower sensitivity as well as a lower specificity in study groups with a greater prevalence¹. The opposite effect has also been reported. A study of Magnetic Resonance Imaging to diagnose multiple sclerosis, reported both a higher sensitivity as well as a higher specificity in the study group with a greater prevalence². On the other hand, when the general results of this study are compared with those of another study with greater prevalence of multiple sclerosis, the latter study reported a lower sensitivity³. Lachs and colleagues⁴, studied dipstick tests in patients suspected of urinary tract infection and found a higher sensitivity and a lower specificity with greater prevalence.

Greater prevalence can be associated with both higher as well as lower sensitivity and specificity. In this paper, we explain some of the underlying mechanisms that can lead to changes in both disease prevalence and in diagnostic accuracy (see Figure 5.1). Prevalence variability itself, as well as the study characteristics that cause prevalence differences, can result either in clinical or artefactual variation

Table 5.1: Characteristics of examples; sens and spec with increasing prevalence

First Author (year)	Target Condition	Index Test	Reference Standard	Prevalence	Sensitivity	Specificity
Flicker (1997)	Dementia	Checklists	Consensus diagnosis	41% vs. 72%	78 to 73 (↓)	88 to 71 (↓)
O'Connor (1996)	MS	MRI	Expert panel	'higher probability'	20 to 70% (↑)	80 to 93% (↑)
Lee (1991)	MS	MRI	Clinical follow-up	43% vs. 53%	84 to 58% (↓)	63 to 91% (↑)
Lachs (1992)	Urinary tract infection	Dipstick	Culture	7% vs. 50%	56 to 92% (↑)	78 to 42% (↓)

Abbreviations: MS=Multiple Sclerosis; MRI=Magnetic Resonance Imaging; ↓ = lower; ↑ = higher.

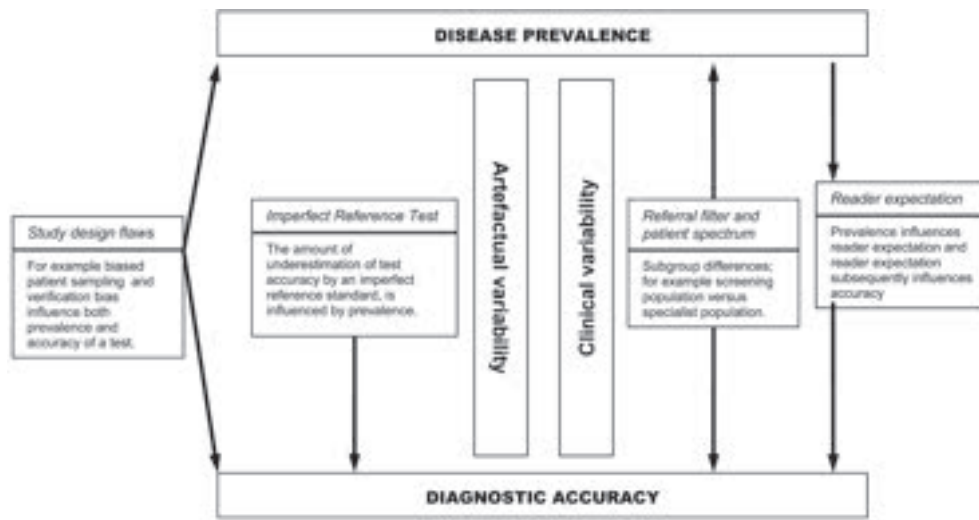


Figure 5.1. Diagram of influences on disease prevalence and diagnostic test accuracy.

in test accuracy, the latter being a consequence of imperfections in study design or execution.

We will first discuss clinical variation, then artefactual variation. In the closing section we will provide guidance for both readers and researchers about how to deal with prevalence differences in study populations and the translation into practice.

5.2 Clinical variability in prevalence and test accuracy

Clinical variability refers to diagnostic test accuracy varying with prevalence because of differences in the patients or the characteristics of the setting in which those patients are being assessed.

5.2.1 Patient Spectrum

Both disease prevalence and test accuracy may be associated with patient spectrum, a term that denotes the severity of disease or the range of comorbidities in the patients studied⁵. Flicker and colleagues studied the diagnostic accuracy of different checklists for dementia in two different settings: a screening group that consisted of elderly people, with memory difficulties, from the general population, with a dementia prevalence of 41%, and a diagnostic care group that consisted of people who were already more or less mentally disabled, with a dementia prevalence of 72%. It is likely that distinguishing patients with dementia from those without

dementia was more difficult in the diagnostic care setting. The more the underlying conditions in these patients look alike, the more false positive results as well as false negative results will be encountered. This is reflected by the lower sensitivity (73% versus 78%) and the lower specificity (71% versus 88%) in the diagnostic care group.

Not only comorbidities affect a test's ability to distinguish people with the target condition from those without this target condition. Many target conditions represent an underlying continuum, ranging from 'barely present' to 'clearly present'. It is possible that the shape of the distribution of the underlying continuum varies with disease prevalence. For example, in situations where disease is common, the distribution may be skewed towards the 'clearly present' end of the spectrum, and sensitivity is higher.

Weiner and colleagues reported on the diagnostic accuracy of an exercise test for coronary artery disease in the coronary artery surgery study⁶. Patients in this study were divided into three groups, based on their symptoms: definite angina, probable angina or nonischemic pain. The prevalence of coronary artery disease was 89% in males with definite angina, 70% in males with probable angina and 22% in males with nonischemic pain. The higher the chances that the symptoms are a manifestation of coronary artery disease, the higher the likelihood that the coronary artery disease is severe, and that a person can be correctly identified as having coronary artery disease. In definite angina, sensitivity will be higher. On the other hand, it will be more difficult to correctly classify persons with definite angina as not having coronary artery disease, so specificity can be expected to be lower. The sensitivity in the group of men who displayed definite angina was 85% while their specificity was 67%, in the group with probable angina sensitivity was 75% and specificity was 74%, and in the group with nonischemic pain, the sensitivity was 54% and specificity 76%.

5.2.2 Referral Filter

Differences in patient spectrum may be caused by differences in study population and clinical setting, but also by prior testing of patients before they are enrolled in the study. Possible effects of prior testing of patients were nicely shown in two studies on the diagnostic accuracy of a diagnostic protocol for children suspected of having appendicitis^{7,8}.

Kosloske and colleagues reported a relatively high sensitivity (99%) in their study, compared to other appendicitis studies⁷. This coincided with a greater prevalence (59%). Although Kosloske claimed that prior testing of children more likely to have appendicitis did not affect sensitivity and specificity, Swarr and Keren pointed out in a comment that prescreening and the related prevalence change had influenced the diagnostic accuracy estimates in this appendicitis study⁹. The severity of the appendicitis is associated with the displayed symptoms, and with the ability of either a general practitioner or an emergency doctor to correctly refer only those

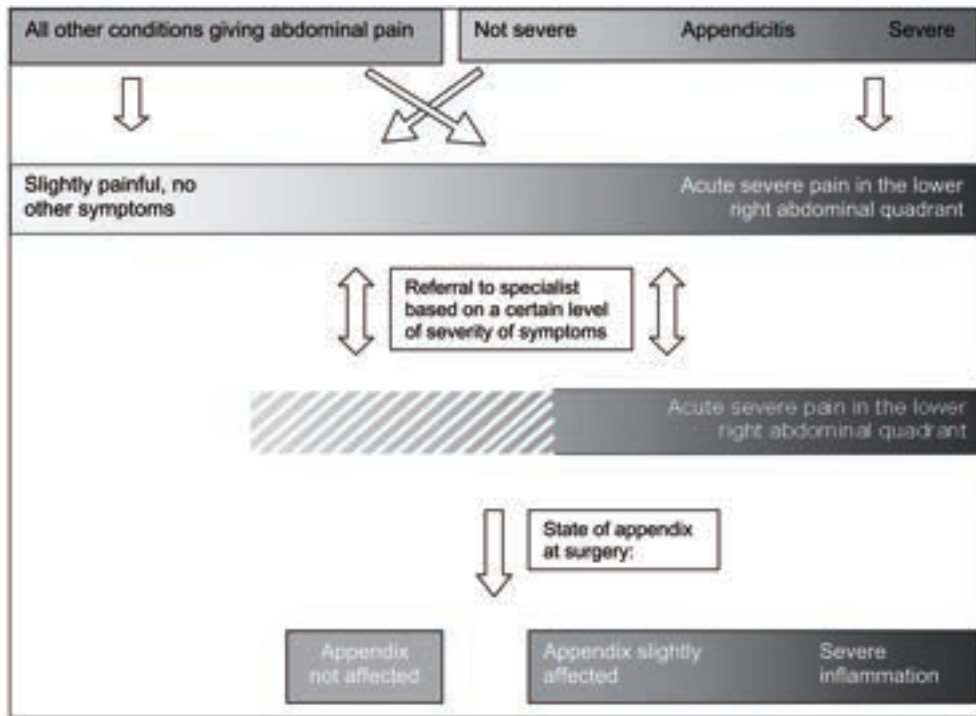


Figure 5.2. Referral of children having abdominal pain.

This diagram shows the relationship between the underlying condition, its symptoms and referral pattern. The top row reflects all children with abdominal pain who present to the general practitioner or to an emergency department. The children will show different levels of pain, but more pain will not always mean more severe appendicitis: some other conditions may be just as painful. Referral to a specialist (and to a study) may be based on symptoms, e.g. the severity of pain or localization of pain. A study that includes only children with severe pain in the lower right abdominal quadrant may end up with a different patient spectrum than a study that includes all children with moderate and severe pain, disregard the localization. The bottom row reflects verification by surgery. The proportion of children that in the end indeed have appendicitis will also differ according to referral pattern.

children that really have appendicitis (see Figure 5.2). Children that are more likely to have appendicitis may express clearer symptoms, resulting in a higher sensitivity. On the other hand, Peña and colleagues sent the children that were most likely to have appendicitis directly to surgery and did not include them to study the accuracy of the diagnostic protocol⁸. They therefore reported a lower prevalence than Kosloske et al.: 36%. This may have resulted in less clear appendicitis cases in the study group, thus leading to more false negatives and a lower sensitivity (94%).

In the above example, prior testing of children led to differences in patient spectrum and thus to variation in prevalence and diagnostic accuracy. More generally, the referral filter is the diagnostic pathway that determines which patients will be referred to the setting where they will be enrolled in the study. Even without result-

ing in apparent differences in patient spectrum, a different referral filter can lead to changes in prevalence and in diagnostic test accuracy.

In a study on the diagnostic accuracy of ultrasound for the detection of breast cancer in young symptomatic women, using mammography as referral filter may affect the prevalence of breast cancer in the study group as well as the diagnostic accuracy of ultrasound. Tables 5.2a and 5.2b show the effect of mammography as a referral filter when assessing the accuracy of ultrasound. Test accuracies derived from a study by Houssami et al. have been applied to a hypothetical population with a breast cancer prevalence of 9%¹⁰. The sensitivity of ultrasound is 82% (82/100) in the overall population. If only mammography positive women are referred for ultrasound, the prevalence increases to 39%. The sensitivity of ultrasound in the mammography-positive group (62/76: 82%) is identical to the overall sensitivity before referral. This is because the errors of mammography and ultrasound in detecting disease are not associated. However, the errors of mammography and ultrasound in declaring breast cancer to be absent are correlated. Hence, the specificity of ultrasound in the overall group differs markedly (880/1000: 88%) from that in women who were positive on mammography (90/120: 75%). With positive mammography as the referral filter, the specificity of ultrasound is lower. These correlated errors may well occur for other reasons than spectrum differences.

5.2.3 Reader Expectation

In 1990, Gianrossi and colleagues reported a meta-analysis on cardiac fluoroscopy to diagnose coronary artery disease¹¹. They found a lower sensitivity in studies with greater prevalence, without any apparent reason for this difference. They reasoned that clinicians who were used to a lower disease prevalence than the prevalence in this study population, were less likely to indicate a patient as having coronary artery disease based on fluoroscopy abnormalities.

Variation in prevalence can be a cause of accuracy differences when it influences the implicit threshold that clinicians use when they judge for example radiological images. This is called reader expectation and may be expected when clinicians switch to a setting with a different prevalence than they were used to. In response to

Tables 5.2a and 5.2b: The effect of using mammography as a referral filter when assessing the accuracy of ultrasound for the diagnosis of breast cancer.

5.2a Breast Cancer				5.2b No Breast Cancer			
	US+	US-	Totals		US+	US-	Totals
M+	62	14	76	M+	30	90	120
M-	20	4	24	M-	90	790	880
Totals	82	18	100	Totals	120	880	1000

Table 5.2a displays the results for women having breast cancer. Table 2b displays the results for women having no breast cancer. US = ultrasound. M=mammography.

prevalence, the clinicians may alter their threshold for declaring a perceived characteristic as abnormal. Compared to those who work in screening, physicians who are more involved in diagnostic examinations (and less involved in screening) may expect higher underlying rates of cancer when reading screening mammograms. This would lead to a lower false negative rate and a higher false positive rate, thus leading to a lower sensitivity and a higher specificity in reading¹².

5.3 Artefactual variability in prevalence and test accuracy

Artefactual variability refers to changes in prevalence due to imperfections in the design and execution of a study. Crucial study design features that relate to both prevalence and diagnostic accuracy are a distorted inclusion of participants in the study and misclassification in the reference standard used for verification of the index test results.

5.3.1 Distorted Inclusion of Participants

Ideally a diagnostic accuracy study includes all patients within a specific period who are suspected of having the target condition and in whom using the test would be considered (consecutive enrollment). This will result in a patient spectrum that reflects as much as possible the range of patients that a clinician will see in practice. Distortion of this ideal inclusion pattern may artefactually affect the prevalence of the target condition in the study group as well as the accuracy of the diagnostic test under study.

The most extreme form of distorted patient inclusion occurs when persons who have the target condition are sampled from a completely different population than the persons who do not have the target condition. Such a design approach, often called a case-control design, can be used without bias if there is appropriate sampling¹⁰. On the other hand, these two-gate designs can be a serious source of bias when cases and controls are sampled from two different populations¹³.

Medeiros and colleagues demonstrated the effects of a two-gate design in a study on the diagnostic accuracy of several tests to diagnose glaucoma¹⁴. They compared two different methods of patient recruitment: the first method comprised consecutive enrollment of patients and the other method was a two-gate design. With the two-gate design, sensitivity was calculated in a group of relatively severe glaucoma patients, while specificity was calculated in a group of healthy volunteers. With the consecutive enrollment design a study group was assembled consisting of patients all suspected of having glaucoma. The Glaucoma Probability Score, which is an automated device to detect glaucomatous damage, had a higher sensitivity and spe-

cificity with the two-gate design (sensitivity of 64% and specificity of 95%) than in the consecutive enrollment design (sensitivity of 35% and specificity of 86%).

In a study with the two-gate design, the researchers selected the cases and the controls themselves and they determined the apparent prevalence. In such an example, the prevalence in the study group is determined by the study design and may not reflect the population prevalence as seen in practice. In other studies, the effect of distorted selection may be more subtle and the prevalence in the study group may seem to reflect the prevalence as seen in practice.

The latter is demonstrated in a study of ultrasound for diagnosing epididymitis with retrospective selection of patients¹⁵. Four different strategies to select patients with epididymitis from existing data files resulted in prevalences ranging from 23% when the broadest selection method was used to 76% with the narrowest selection method. With greater prevalence, sensitivity decreased from 83% to 76% and specificity decreased from 97 to 79%.

If more conditions were included, such as testicular torsion, orchitis or testicular carcinoma, prevalence was lower and it was easier for the readers of the ultrasound images to differentiate between patients who had epididymitis and patients who had another scrotal disease. When they only looked at patients with epididymitis or epididymo-orchitis in the differential diagnosis, the prevalence of epididymitis was greater but it also became more difficult to differentiate between patients with and without epididymitis.

Exclusion of patients can also have effects in the opposite direction. By excluding related target conditions that challenge a test's ability to detect the target condition as well as the ability to identify the patients without the target condition, a test's sensitivity and specificity will be higher but so will prevalence. By excluding these related conditions from the study or the subsequent analyses, a test may seem to perform better, a phenomenon known as limited challenge¹⁶. Note that what will be called limited challenge in one situation may be called a difference in patient spectrum in another situation. Excluding obese patients in a study on the accuracy of ultrasound can be regarded as an example of limited challenge, as it is more difficult to distinguish abnormalities by ultrasound in obese patients than in non-obese patients. On the other hand, if the ultrasound is not used in obese patients, the exclusion of obese patients and the resulting diagnostic accuracy of the ultrasound will fairly reflect the clinical situation.

5.3.2 Verification Bias

Diagnostic test accuracy is assessed by verifying the results of the test under evaluation with the result of a single reference standard in every patient in the study. Verification bias occurs if not all patients are verified, or if some patients are verified by a second or a third reference standard.

An example of verification bias can be found in the meta-analysis of Mol and colleagues¹⁷. They assessed the accuracy of nuchal translucency measurement for Down syndrome detection. Some studies used two reference standards: fetal karyotyping in fetuses with an increased nuchal translucency, whereas pregnancy outcome was awaited in fetuses that showed a normal measurement. In the studies with such a verification bias, the prevalence ranged from 0.1 to 0.9% (pooled estimate 0.4%). In studies without verification bias, the prevalence ranged from 0.2 to 2.3% (pooled estimate 1.1%). The pooled sensitivity in the last group of studies was found to be lower (55%) than that in the studies with verification bias (77%), whereas specificity remained unaffected (96% and 97%).

In these studies, test positives were verified with another reference standard than the test negatives. This is called differential verification: some participants receive a different reference standard, conditional on the index test result. The reference standard in the test negatives was follow-up. Between testing and birth, fetuses may be aborted (and thus be excluded from the analysis) or Down syndrome may not have been recognized directly at birth. In general, the effects of differential verification very much depend on the uniformity in the characteristics of all used reference standards.

Another form of verification bias is partial verification. Partial verification occurs when not all participants receive the reference standard. The effects on prevalence and diagnostic accuracy depend on whether the reference standard was randomly allocated to patients or not. In the majority of studies with partial verification, most index test positive cases are verified, whereas index test negative cases are likely to be verified only in case of an increased pretest suspicion. In that case, a number of false negatives are not detected as such, and even more true negatives drop out of the analysis. The result is an overestimation of prevalence and of sensitivity.

Based on a systematic review of sources of bias in accuracy studies, Whiting et al.¹⁶ reported that differential verification led to overestimation of overall accuracy and that partial verification always led to overestimation of sensitivity but not always on specificity.

5.3.3 Reference standard misclassification

Another artefactual modifier of prevalence is the use of an imperfect reference standard. Because a perfect reference standard is unlikely, reference standard imperfections will play a role in most accuracy studies and this effect may have been present in all examples mentioned so far. The study of Lachs⁴, for example, resulted in a letter from Boyko¹⁸, who raised the possibility that the spectrum bias Lachs and colleagues described was partly due to the imperfect reference test they used. The same issue was mentioned by Evans¹⁹, some ten years later, in response to another article on spectrum variability²⁰.

The issue of reference standard misclassification had already been raised in 1966 by Buck and Gart^{21,22}. Using a hypothetical example, they showed that in the presence of an imperfect reference standard, the reported accuracy will always be lower than the true accuracy, but sensitivity will increase towards its original value as prevalence increases, while specificity decreases. The same was also described by Brenner and Gefeller in 1997 and by Miller in 1998^{23,24}. In addition, imperfect reference standards also bias the reported prevalence. The following example illustrates this effect.

Let us assume that the prevalence of pulmonary embolism in hospitalized patients is 10% and that the sensitivity of a D-dimer test is 95% and its specificity 60%. In a study with 1000 patients, this may lead to 455 patients with a positive index test and 545 patients with a negative index test. In a study where patients were verified by a ventilation perfusion scan, with a sensitivity of 95% and a specificity of 90%, the estimated prevalence of pulmonary embolism will be around 18.5%. The estimated sensitivity will then be 68% and specificity 60%. Because estimates of both prevalence and accuracy are affected by reference standard misclassification, accuracy may artefactually seem to vary among subgroups of patients with different prevalences²⁵.

5.4 Conclusions

By their mathematical definition, sensitivity and specificity do not depend on the disease prevalence. Yet we have shown a series of examples that prevalence and diagnostic test accuracy may covary with prevalence. These examples were from both systematic reviews, which showed variation between studies, and from individual studies, which showed variation between patient subgroups. The parallel variability of prevalence and accuracy can occur through clinical mechanisms, such as patient spectrum, referral filter, or reader expectation, and artefactual mechanisms, which include distorted inclusion of participants, verification bias and reference standard misclassification. An awareness of these mechanisms and the way they can affect diagnostic accuracy is essential for a balanced translation of study results into clinical practice.

In clinical practice, Bayes' rule is often applied, in which the likelihood ratio of a test is used to translate the pre-test probability to post-test probability. The pre-test probability is often based on the disease prevalence. The likelihood ratio of a test is a function of the test's sensitivity and specificity and is not a fixed test property. A likelihood ratio calculated from a study with a prevalence of 5% can therefore not be blindly used to calculate the post-test probability in a population with a prevalence of 20%.

The latter was demonstrated in the study of Van der Schouw and colleagues on diagnostic accuracy of ultrasonography for epididymitis mentioned earlier¹⁵. In the patient group that had clearly epididymitis or epididymo-orchitis in the differential diagnosis, prevalence of epididymitis was 81%, post-test probability was 94% and the positive likelihood ratio was 4. In the patient group with diseases mimicking epididymitis in the differential diagnosis, the prevalence of epididymitis was 39%, the post-test probability was 91% and the positive likelihood ratio was 16. If we would have had only the results of the group with a prevalence of 81% (and a positive likelihood ratio of 4) and applied those to the group with a prevalence of 39% by using Bayes' rule and the positive likelihood ratio of 4, we would have estimated a post-test probability of 72%. The latter differs markedly from the actual post-test probability in that group, which is 91%.

Other authors have also emphasized the relation between study characteristics and changes in diagnostic test accuracy²⁶⁻²⁸. Unfortunately, study design features and characteristics of the population or referral filter are still badly reported²⁹. Prevalence is therefore the most apparent key feature of studies. The examples and mechanisms in this paper illustrate how prevalence can be used to signal study design deficiencies and crucial differences in patient characteristics. Hopefully a more widespread dissemination and implementation of the Standards for the Reporting of Diagnostic accuracy studies (STARD) by authors and journals will enable readers to signal study characteristics directly³⁰.

Clinicians who use the diagnostic literature in their daily practice should carefully define their clinical question first: in what population is the test going to be used, what is the clinical setting, and what is the referral filter. Studies not addressing that question and studies with obviously improper designs, such as those relying on comparisons between healthy controls and severely diseased, are unlikely to be helpful and may not be considered further³¹.

Studies being considered further can be expected to show variability in test accuracy. By examining how accuracy varies with prevalence, an understanding of more subtle biases and sources of between-study variability in accuracy can be of help. Reasons for artefactual variability, as discussed in this paper, should be identified first. Were certain patient groups excluded from the study; was the same reference standard used in all patients? Although their effect on prevalence may vary, as seen in the examples we used in this paper, both limited challenge and verification will most often lead to higher diagnostic accuracy^{32,33}. The magnitude of the overestimation due to flaws in study design will vary. The severity of this overestimation will depend on clinical question and the decision that has to be made. The effect of reference standard misclassification will lead to more predictable changes in test accuracy. In case of an imperfect reference standard, sensitivity will be less underestimated and specificity will be more underestimated with greater prevalence.

When the reader is ensured of the absence of bias, reasons for clinical variability, such as differences in patient groups, have to be identified. The central question here is: do the patients in this study reflect my clinical population? The combination of setting, referral filter and prevalence can be used to select those studies that are most appropriate for the clinical question. If referral filter and setting are badly reported, prevalence can serve as a guiding tool: does the prevalence of the studied population reflect my own patient population?

When both clinical and artefactual mechanisms, with possibly conflicting effects are present, the net result may be difficult to predict. Systematic reviews of diagnostic accuracy studies that take variability in prevalence into account, may throw some more light on these mechanisms and their effects.

Sensitivity and specificity are not fixed test characteristics, but test properties, that describe the behaviour of the test in a particular situation. As the setting, filter, or patient group changes, prevalence and accuracy may change. For this reason, variation in disease prevalence and test accuracy between studies can act as a flag for clinicians to detect important differences in study population or study design, affecting accuracy.

References

1. Flicker L, Logiudice D, Carlin JB, Ames D. The predictive value of dementia screening instruments in clinical populations. *Int J Geriatr Psychiatry*. 1997; 12(2):203–9.
2. O'Connor PW, Tansay CM, Detsky AS, Mushlin AI, Kucharczyk W. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. *Neurology*. 1996; 47(1):140–4.
3. Lee KH, Hashimoto SA, Hooge JP, Kastrukoff LF, Oger JJ, Li DK, Paty DW. Magnetic resonance imaging of the head in the diagnosis of multiple sclerosis: a prospective 2-year follow-up with comparison of clinical evaluation, evoked potentials, oligoclonal banding, and CT. *Neurology*. 1991; 41(5):657–60.
4. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med*. 1992; 117(2):135–40.
5. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978; 299(17):926–30.
6. Weiner DA, Ryan TJ, McCabe CH, Kennedy JW, Schloss M, Tristani F, Chaitman BR, Fisher LD. Exercise stress testing. Correlations among history of angina, ST-segment response and prevalence of coronary-artery disease in the Coronary Artery Surgery Study (CASS). *N Engl J Med*. 1979; 301(5):230–5.
7. Kosloske AM, Love CL, Rohrer JE, Goldthorn JF, Lacey SR. The diagnosis of appendicitis in children: outcomes of a strategy based on pediatric surgical evaluation. *Pediatrics*. 2004; 113(1 Pt 1):29–34.
8. Garcia Pena BM, Mandl KD, Kraus SJ, Fischer AC, Fleisher GR, Lund DP, Taylor GA. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. *JAMA*. 1999; 282(11):1041–6.
9. Swarr D and Keren R. Comparison of alternative diagnostic approaches for managing appendicitis in children: the effect of disease prevalence and spectrum. *Pediatrics* 2004, 114(2):513–4.
10. Houssami N, Irwig L, Simpson JM, McKessar M, Blome S, Noakes J. Sydney Breast Imaging Accuracy Study: Comparative sensitivity and specificity of mammography and sonography in young women with symptoms. *Am J Roentgenol*. 2003; 180(4):935–40.
11. Gianrossi R, Detrano R, Colombo A, Froelicher V. Cardiac fluoroscopy for the diagnosis of coronary artery disease: a meta analytic review. *Am Heart J*. 1990; 120(5):1179–88.
12. Smith-Bindman R, Chu P, Miglioretti DL, Quale C, Rosenberg RD, Cutter G, Geller B, Bacchetti P, Sickles EA, Kerlikowske K. Physician predictors of mammographic accuracy. *J Natl Cancer Inst*. 2005; 97(5):358–67.
13. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem*. 2005; 51(8):1335–41.
14. Medeiros FA, Ng D, Zangwill LM, Sample PA, Bowd C, Weinreb RN. The effects of study design and spectrum bias on the evaluation of diagnostic accuracy of confocal scanning laser ophthalmoscopy in glaucoma. *Invest Ophthalmol Vis Sci*. 2007; 48(1):214–22.
15. Van der Schouw YT, Van Dijk R, Verbeek AL. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. *J Clin Epidemiol*. 1995; 48(3):417–22.
16. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med*. 2004; 140(3):189–202.

17. Mol BW, Lijmer JG, van der Meulen J, Pajkrt E, Bilardo CM, Bossuyt PM. Effect of study design on the association between nuchal translucency measurement and Down syndrome. *Obstet Gynecol.* 1999; 94(5 Pt 2):864–9.
18. Boyko EJ. Leukocyte esterase tests detect pyuria, not bacteriuria. *Ann Intern Med.* 1993; 118(3):230.
19. Evans AT. Subgroup variation in diagnostic test evaluation. *Ann Intern Med.* 2003; 138(8):686.
20. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med.* 2002; 137(7):598–602.
21. Buck AA, Gart JJ. Comparison of a screening test and a reference test in epidemiologic studies. I. Indices of agreement and their relation to prevalence. *Am J Epidemiol.* 1966; 83(3):586–92.
22. Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. *Am J Epidemiol.* 1966; 83(3):593–602.
23. Brenner H and Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med.* 1997; 16(9):981–91.
24. Miller WC. Bias in discrepant analysis: when two wrongs don't make a right. *J Clin Epidemiol.* 1998; 51(3):219–31.
25. Biesheuvel C, Irwig L, Bossuyt P. Observed Differences in Diagnostic Test Accuracy between Patient Subgroups: Is It Real or Due to Reference Standard Misclassification? *Clin Chem.* 2007; 53(10):1725–9.
26. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology.* 1997; 8(1):12–7.
27. Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. *Am J Med.* 1984; 77(1):64–71.
28. Knottnerus JA and Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol.* 1992; 45(10):1143–54.
29. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, Bouter LM, de Vet HC. The quality of diagnostic accuracy studies since the STARD statement: has it improved? *Neurology.* 2006; 67(5):792–7.
30. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med.* 2003; 138(1):40–4.
31. Sackett DL and Haynes RB. The architecture of diagnostic research. *BMJ.* 2002; 324(7336):539–41.
32. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 1999; 282(11):1061–6.
33. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ.* 2006; 174(4):469–76.





**Accuracy of fibronectin tests for the prediction of
pre-eclampsia: a systematic review**

**Mariska M.G. Leeflang, Jeltsje S. Cnossen, Joris A. van der Post,
Ben Willem Mol, Khalid S. Khan, Gerben ter Riet**

Eur J Obstet Gynecol Reprod Biol. 2007;133(1):12-9

Abstract

Background: The purpose of this study was to review systematically all studies that assessed the accuracy of maternal plasma fibronectin as a serum marker for early prediction of pre-eclampsia.

Methods: We therefore assessed studies that reported on fibronectin as serum marker for pre-eclampsia before the 25th gestational week. For the selected studies, sensitivity and specificity were calculated and plotted in ROC-space.

Results: We included twelve studies, of which only five studies reported sufficient data to calculate accuracy estimates, such as sensitivity and specificity. These five studies reported on 573 pregnant women of whom 109 developed pre-eclampsia. At a sensitivity of at least 50%, specificities ranged between 72 and 96% for cellular fibronectin. For total fibronectin, these numbers were 42 to 94%.

Conclusions: Fibronectin seems to be a promising marker for the prediction of pre-eclampsia. However, further studies are needed to determine whether the accuracy of this test is sufficient to be clinically relevant.

6.1 Background

Pre-eclampsia is among the largest single causes of maternal and foetal mortality and morbidity world wide¹⁻³. It has a long pre-clinical phase before signs become clinically manifest during the second half of pregnancy. Good prediction will enable to redirect intensified prenatal care from all pregnant women to those women and foetuses who are at higher risk, and to more effectively evaluate interventions for prevention of pre-eclampsia⁴⁻⁸. Also, women at high risk could benefit from increased surveillance, preventive therapies like aspirin and early diagnosis^{9,10}.

Maternal and perinatal mortality and morbidity result from maternal organ failure, foetal growth restriction and premature delivery. Maternal endothelial damage and inadequate placental development are both involved in the genesis of pre-eclampsia¹¹. Therefore, a number of products released from the placenta and biochemical markers for endothelial damage were tested for their ability to predict the onset of pre-eclampsia. One of these possible markers was fibronectin (Fn), a glycoprotein that plays a role in a variety of biological functions.

Several subtypes of Fn exist. Inflammation, vascular injury and malignancy are generally associated with increased expression of the ED-A (also called ED-1+ or oncofoetal Fn) and ED-B (also called ED-2+) forms of Fn, particularly in the blood vessel walls¹²⁻¹⁴. ED-A (oncofoetal) Fn is also released by the placenta and has been used as a predictor for preterm birth^{15,16}.

Several studies showed that, on average, women destined to develop pre-eclampsia had higher plasma Fn concentrations than (pregnant) controls. However, these studies differ in, for example the type of test that is evaluated, the study population, and scientific rigour. Earlier reviews about the prediction of pre-eclampsia that also included Fn measurements reported conflicting results or did not differentiate between ED-A or ED-B Fn (only 5% of all Fn in plasma) and total Fn (all subtypes of Fn)¹⁷⁻²⁶. The most recent review reported low predictive accuracy of the Fn tests¹⁸. However, this was based on only one study. In addition, this review has been criticized for performing the crucial steps of screening of bibliographies and data-extraction using a single reviewer only and suboptimal statistical methods²⁷.

We conducted a systematic review of the available evidence to obtain valid and reliable estimates of predictive accuracy of Fn assays for the early (< 25th gestational week) prediction of pre-eclampsia.

6.2 Methods

6.2.1 Study selection and data extraction procedures

We developed an electronic search strategy for the general databases: MEDLINE (1953-2004), and EMBASE (1980-2004), and specialist databases: The Cochrane Library (2004:3), and MEDION (1974-2004; www.mediondatabase.nl). This search was updated in April, 2006. The search strategy consisted of MeSH and keyword terms related to pre-eclampsia combined with methodological filters for identification of diagnostic test and aetiological studies^{28,29}. Reference lists of review articles and eligible primary studies were checked to identify cited articles not captured by electronic searches. The electronic search strategy is available from the authors.

Studies were selected in a three-stage process. First, titles and/or abstracts of all references (Reference Manager 10.0) were scrutinized by one reviewer for studies that reported on any test used in predicting pre-eclampsia (JC, GtR, JvdP and BWM). Then, for this particular review, a second reviewer scrutinized all references with “fibronectin” as keyword or as word in title or abstract to ensure independent duplicate selection (JC). Final in-/ exclusion decisions were made after independent duplicate examination of the full manuscripts of selected references (JvdP and ML). Studies were included if they reported on Fn testing in maternal serum or plasma before the 25th gestational week (mean). Language restrictions were not applied. Any disagreements were resolved by consensus and, if necessary, by a third reviewer (JC). For each included article, data on study characteristics (both clinical and methodological) and on test accuracy were extracted independently by two reviewers (JvdP and ML) on piloted data extraction forms. Disagreements were resolved by consensus. Study characteristics consisted of women’s risk classifications, characteristics of the index test and the reference standard.

6.2.2 Quality assessment

The methodological quality of the selected primary studies was assessed using pre-defined criteria based on elements of study design, conduct and analysis which are likely to have a direct relationship to bias in a test accuracy study³⁰⁻³². For this purpose, we used the QUADAS list³³, a tool for quality assessment of diagnostic accuracy studies. This checklist was adapted with respect to timing of the test, patient spectrum (some patient characteristics, such as being normotensive and non-proteinuric, are part of the reference standard), partial verification and the index test being part of the reference standard. We also assessed the occurrence of a potential treatment paradox (mainly the use of antihypertensive drugs; yes or no), because this review deals with prediction instead of diagnosis. Patient spectrum was judged representative for general pregnant populations when eligible women were consecutively recruited and the incidence of pre-eclampsia did not exceed 4%.

6.2.3 Data synthesis: main analysis

For each study, we constructed a 2-by-2 table cross-classifying Fn results and the occurrence of pre-eclampsia. Sensitivity, specificity and likelihood ratios were calculated. We assessed the heterogeneity of results between studies looking at the distribution of sensitivities and specificities in the receiver operating characteristic (ROC) plot. Because of the differences in study characteristics, we considered meta-analysis to generate summary estimates not appropriate.

6.3 Results

6.3.1 Included studies

Figure 6.1 summarizes the selection process for studies on Fn and prediction of pre-eclampsia. Twelve studies³⁴⁻⁴⁵ met the inclusion criteria, eight cohort studies³⁴⁻⁴¹ and four nested case control studies⁴²⁻⁴⁵ (Table 6.1). All case control studies selected incident cases of pre-eclampsia and non pre-eclamptic controls. Three were matched case control studies^{42,44,45}, matching occurred on factors such as maternal and gestational age. No studies classified the cases into severe and mild pre-eclampsia. The cohort studies were all conducted in hospitals providing secondary or tertiary

Figure 6.1. Study selection process for this review.
Of the finally included 12 primary studies, five reported sufficient data for 2x2 tables.

Table 6.1. Key characteristics of included studies.

First Author (Year)	Setting (no. centres)	Country	Design	n**	Incidence of PE (%)	Inclusion Criteria	Exclusion Criteria	Reference Standard	Fn fraction measured
Lockwood ²² (1990)	Primary care (1)	USA	nested and matched CC	57	4.50	Singleton pregnancies, normotensive <20 wks gest; not all were primigravids.	IDDM, CH, abruptio placentae and infections, history of previous PE	BP 140/90 mmHg, rise in systolic or diastolic BP of 30 resp. 15 mmHg (in seated position, Korotkoff phase V); 1 gm/L proteinuria; at least two occasions >6 h apart.	total Fn and ED-A Fn
Taylor ²³ (1991)	Mixed settings (2)	USA	nested and matched CC	38	NR	Normotensive and non-proteinuric <20 wks gest.	Identification of any chronic metabolic disease, evidence of illicit drug use or the failure of elevated BP, hyperuricemia or proteinuria to resolve within 12 weeks after delivery.	Rise in systolic or diastolic BP of 30 resp. 15 mmHg (in seated position, Korotkoff phase V); proteinuria ≥ 0.5 gm/24 h or ≥ 30 mg/dl in a catheterized specimen; hyperuricemia.	ED-B Fn
Friedman ²⁴ (1992)	Secondary / tertiary care (1)	USA*	nested CC	20	NR	NR	NR	BP 140/90 mmHg, rise in systolic or diastolic BP of 30 resp. 15 mmHg; proteinuria $\geq 1+$ (categorized or $\geq 2+$ voided); and hyperuricemia.	Fetal Fn
Mulligan ²⁵ (1994)	Mixed settings (NR)	Ireland*	cohort	36	11.1	Primigravids.	NR	Referred to Davey and McGillivray, 1988.	Total plasma Fn
Jones ²⁶ (1996)	Secondary / tertiary care (1)	Australia	cohort	171	19.3	Singleton pregnancies, normotensive <20 wks gest.	NR	Referred to Beischer & Mackay	Serum Fn
Soltan ²⁷ (1996)	Secondary / tertiary care (1)	Egypt	cohort	88	NR	Normotensive and non-proteinuric <20 wks gest.	IDDM, CHD, history of cardiovascular or renal disease, aspirin therapy, antiprostaglandins, calcium, albuminuria, any abnormality.	Rise in systolic or diastolic BP of 30 resp. 15 mmHg (in seated position, Korotkoff phase V); or ≥ 300 mg proteinuria in 24 h; or generalized edema with one of the above.	Plasma Fn
Pluurberg ²⁸ (1996)	Secondary / tertiary care (1)	The Netherlands	cohort	228	7.70	Singleton pregnancies, normotensive <20 wks gest.	IDDM, CHD, APS, age < 18; miscarriage before 16 wks; treatment: Crohn's, idiopathic hypoglycemia, myotonia uteri, uterine anomaly, sickle cell anemia, trisomy-21 infant, twin pregnancy, congenital abnormalities.	Diastolic BP 90 mmHg, rise in diastolic BP of 15 mmHg (in seated position, Korotkoff phase V); proteinuria of 0.3 g in 24 h.	Total plasma Fn
Sudin ²⁹ (1999)	Secondary / tertiary care (1)	India	cohort	100	14.0	Singleton pregnancies, normotensive <20 wks gest.	DM, multiple pregnancies, history of: trauma, surgery, blood transfusion 6 weeks prior, edema complicating pregnancy and coagulation disorders.	BP 140/90 mmHg on more than two occasions 6 h apart; proteinuria ≥ 0.3 gm/dl in 24 h; $1+$; pedal edema of $1+$ after 12 hours of rest.	Plasma Fn
Islam ³⁰ (2001)	Secondary / tertiary care (1)	Switzerland	cohort	198	4.50	Normotensives and hypertensives; some women were proteinuric < 20 wks gest; not all women were primigravids	No exclusion criteria, comorbidities classified in subgroups	Diastolic BP 90 mmHg on at least two occasions and >0.3 g proteinuria/day.	ED-B Fn
Outspan ³¹ (2001)	Secondary / tertiary care (1)	Sweden*	cohort	228	2.60	Normotensive and non-proteinuric <20 wks gest.	NR	BP $\geq 140/90$ mmHg and albuminuria ≥ 0.3 g/day or $2+$ dipstick.	Total Plasma Fn
Chavarria ³² (2002)	Primary care (1)	Mexico	Nested and matched CC	78	6.88	Normotensive and non-proteinuric <20 wks gest.	IDDM, CHD, APS, SLE, miscarriages, multiple pregnancies, essential hypertension, aspirin therapy, gest/transient hypertension, gest-DM, diphyhydramnios, stillbirth, preterm delivery.	BP $\geq 140/90$ mmHg, rise in systolic or diastolic BP of 30 resp. 15 mmHg (in sitting position, Korotkoff phase V), at least twice, ≥ 6 h apart; >300 mg proteinuria in 24 h or dipstick $1+$; and edema $1+$ after bed rest.	ED-B Fn
Madzaji ³³ (2005)	Secondary / tertiary care (1)	Turkey	cohort	122	11.5	Singleton pregnancies, normotensive and non-proteinuric; not all were primigravids.	NR	BP 140 / 90 mm Hg or greater, 6 h or more apart mmHg (in sitting position, Korotkoff phase V), and consistent proteinuria (300 mg/day or more).	Plasma Fn

In the second column, in parentheses, the numbers of centers in which the study was conducted is stated. * = according to author's affiliation; ** = patients of which fibronectin levels were measured in first or second trimester. CC = case control; NR = not reported; PE = preeclampsia; wks gest = weeks gestation; IDDM = Insulin Dependent Diabetes Mellitus; BP = Blood pressure; Fn = fibronectin; h = hours; pp = post partum.

care, except for one study, which was conducted in an unclear reported number of hospitals providing primary and secondary care³⁴. Since several definitions of pre-eclampsia prevail worldwide, different reference standards were used. Three studies^{39,40,42} included the presence of oedema in their definition of pre-eclampsia and one⁴⁵ included the presence of hyperuricaemia. The incidence of pre-eclampsia varied from 2.6% to 19.3% (median 7.7%), but in three studies the incidence in the studied population could not be extracted. Mean maternal ages varied from 19 to 31 years. Treatment with aspirin, other anti-inflammatory or anti-hypertensive drugs was only reported when treatment was one of the exclusion criteria. Two studies did not report any selection criteria^{34,43}. In general, the reference test was described in sufficient detail, whereas the index test was not. Blind assessment of either the index test or the reference standard was also poorly reported. Figure 6.2 shows the assessed quality items.

6.3.2 Data analysis

Of the 12 studies included in the review, three studies reported the measurement of total plasma Fn^{34,37,38}, four measured cellular Fn^{35,39,42,45} and one study measured both⁴⁴. Although insufficient details were provided by three other studies, the results indicate that four of them measured total plasma Fn^{36,39,41,45}. The twelve stud-

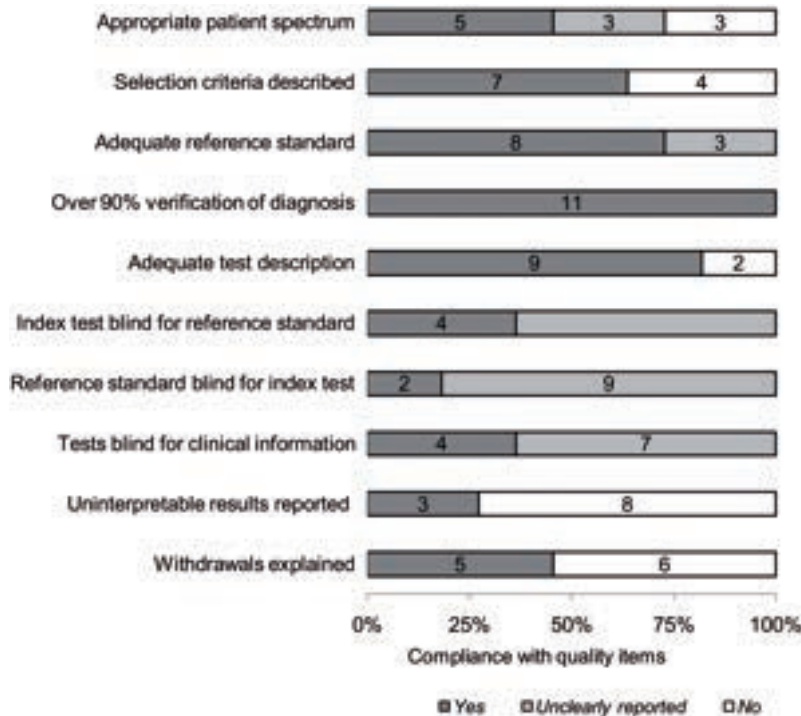


Figure 6.2. Methodological quality of included studies.
Data presented as 100% stacked bars, figures in the stacks represent number of studies.

ies all report assays that are based on immunological principles. Seven studies reported ELISA assays and five of those reported commercially available test kits. Three studies reported other commercially available tests and two reported only the immunologic principle.

Because two authors reported explicitly non-Normal distributions of the Fn-values and one other used non-parametric statistical analyses, we decided not to recalculate Normal distributions from mean and SDs in order to construct 2x2 tables that way. Thus, only five studies reported sufficient details to replicate 2x2 tables and to calculate measures of predictive accuracy^{38,39,41-43}. These studies included a total of 573 pregnant women of whom 109 developed pre-eclampsia. One of those studies, Chavarria et al.⁴², reported ROC-curves separately for Fn values in weeks 18 to 22 and in weeks 22 to 26. However, only for weeks 22 to 26, the results were also reported in a table. When we compared the values of the ROC curve with the values in the table (by labelling the depicted dots with the reported threshold values), the sensitivities reported in the table did not entirely match with the sensitivities reported in the figure. Therefore, the thresholds presented here may slightly differ from the original results. The results are listed in Table 6.2 and Figure 6.3.

Table 6.2 Measures of accuracy.

First Author	Fn fraction	Gest. Period	Threshold (µg/ml)	Sensitivity	Specificity	LR+	LR-
Lockwood ⁴⁴	ED 1+	1st trim	2.8	1.00	0.75	4.00	0.00
			3	0.67	0.75	2.67	0.44
			3.2	0.67	0.75	2.67	0.44
			3.4	0.67	0.75	2.67	0.44
	ED 1+	2nd trim	3.6	0.50	0.88	4.00	0.57
			3.9	0.85	0.74	3.26	0.20
			4.2	0.80	0.78	3.68	0.26
Chavarria ⁴²	ED-B Fn	2nd trim	4.6	0.55	0.83	3.16	0.54
			5	0.50	0.96	11.50	0.52
			3.5	0.74	0.72	2.64	0.36
			3.6	0.70	0.75	2.80	0.40
			3.7	0.64	0.82	3.56	0.44
Lockwood ⁴⁴	Total Fn	1st trim	3.8	0.63	0.85	4.20	0.44
			3.9	0.56	0.88	4.67	0.50
			347	0.83	0.63	2.22	0.27
			370	0.67	0.63	1.78	0.53
Lockwood ⁴⁴	Total Fn	2nd trim	393	0.50	0.75	2.00	0.67
			320	0.70	0.43	1.24	0.69
			350	0.55	0.74	2.11	0.61
Soltan ³⁹	Total Fn	14-24 wks	293.03	0.65	0.94	11.46	0.37
Paarlberg ³⁸	Total Fn	1st trim	240	0.52	0.64	1.47	0.74
		2nd trim	230	0.69	0.67	2.09	0.46
Madazli ⁴¹	Total Fn	21-26 wks	370	0.64	0.86	4.57	0.42

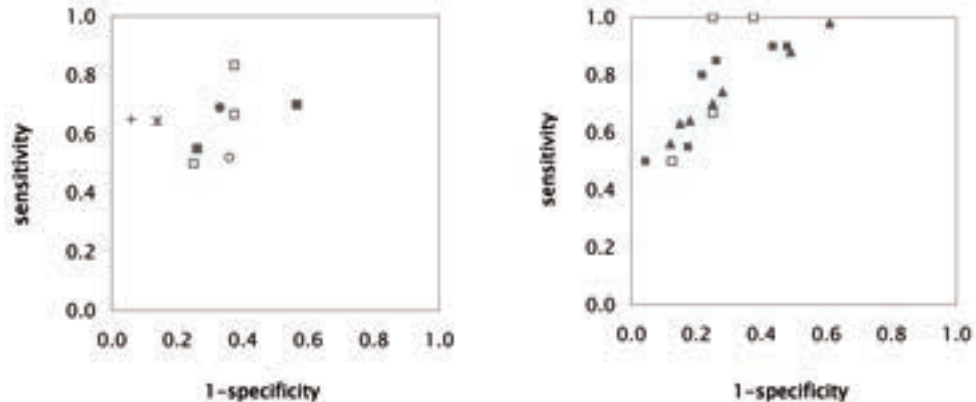


Figure 6.3. ROC plot of the cellular(a) and total(b) Fn assays at various thresholds.

Figure 6.3a. ROC plot of the cellular Fn assays at various thresholds. Depicted are the sensitivities and specificities of Lockwood et al. (\square , first trimester and \blacksquare , second trimester) and Chavarria et al. (\blacktriangle , second trimester). Sensitivities lower than 50% are not depicted. Figure 6.3b. ROC Plot of the total Fn assays. Depicted are the sensitivities and specificities of Lockwood et al. (\square , first trimester and \blacksquare , second trimester), Soltan et al. (+, week 14-24), Paarlberg et al. (\circ , first trimester and \bullet , second trimester) and Madazli et al. (*, week 21-26). The study of Lockwood et al. provided results at various thresholds. Sensitivities lower than 50% are not depicted.

The sensitivities of all Fn assays vary widely, depending on the chosen threshold (Table 6.2). Requiring a sensitivity of at least 50%, the specificity achieved with the cellular Fn assays ranged from 72% to 96%. For the total Fn assays these specificities ranged from 43% to 94%. The positive Likelihood Ratios ranged from 1.64 to 11.5 for the cellular Fn assays and from 1.24 to 10.8 for the total Fn assays. A Likelihood Ratio of 4.67 would increase a pre-test probability to develop pre-eclampsia of 5% to a post-test probability of 20%. The negative Likelihood Ratios varied from 0.0 to 0.57 for the cellular Fn assays and from 0.27 to 0.74 for the total Fn assays. This implies that a negative cellular Fn test result may decrease a pre-test probability of 5% to a post-test probability that approximates 0. Figure 6.3a and 6.3b show the ROC plots. Figure 6.3a only shows the results of the cellular Fn assays. These seem to allow a summary ROC curve. However, these two studies measured different types of cellular Fn (ED-A versus ED-B), assessed first and second trimester and Lockwood et al. did not report on the type of assay used. Therefore, we decided not to draw a summary ROC curve or calculate pooled estimates. Figure 6.3b shows the sensitivities and specificities of the total Fn assays. These studies were also methodologically and clinically heterogeneous; hence we did not calculate pooled estimates here either.

6.4 Discussion

On reviewing 12 studies and analysing five, we found that the accuracy of plasma determination of Fn before the 25th (mean) gestational week to predict pre-eclampsia appears to vary widely among the studies. Because a Normal distribution of Fn-levels could not be assumed, the conclusions are based on only five studies. The exclusion of the other seven studies, that included a total of 791 women, reduced the statistical power of this review. Unfortunately, the extent to which its main conclusions are affected remains speculative. The included studies differed from each other in several aspects, for example, for study design, Fn fraction measured, cut-off values used to determine positive results, incidence of pre-eclampsia, and country where the study was conducted. Furthermore, reference standards (the criteria for pre-eclampsia) varied over the studies as well. None of these five studies reported about blinding of the reference test, whereas the index test is only well described (with manufacturer and inter- and intra-assay variations) by Chavarria and co-workers⁴². These characteristics may artificially inflate or reduce the true sensitivities and specificities^{31,46}. We were unable to analyse the effects of these biases and variations in this review due to the limited number of primary studies yielding usable results. Lockwood et al.'s study⁴⁴ contains some direct evidence that measurement of cellular Fn is more informative than that of total Fn. This study does not indicate that measurement of (cellular) Fn in the 2nd trimester is more useful than in the 1st trimester.

Earlier reviews about the prediction of pre-eclampsia that also included Fn measurements¹⁷⁻²⁶ report conflicting results and did not always differentiate between cellular and total Fn. Conde-Agudelo and colleagues reviewed methods for prediction and screening of pre-eclampsia twice^{17,18}. The conclusion in the first review was based on three studies and in the second review on one study. In addition, this review has been criticized for performing the crucial steps of screening of bibliographies and data-extraction using a single reviewer only and suboptimal statistical methods²⁷.

Because the results of the cellular Fn assays on average seem to have a slightly better performance than the total Fn assays, we think that further research should focus on the use of cellular Fn for the prediction of pre-eclampsia. Such studies should report according to the STARD recommendations for diagnostic accuracy studies⁴⁷. In particular, more details on blinding, concomitant treatment, entry criteria, and the exact Fn technology used is important to readers and reviewers alike. Furthermore, added value of Fn determination given patient information, such as history items, available at the time of assay is an important issue and usually requires multivariable analysis⁵⁰.

At this point, it is not yet possible to advise clinicians on the optimal threshold to achieve a particular specificity in their daily practice. However, this review shows that when both sensitivity and specificity are not allowed to drop below 50%, the

cellular assays can be used to exclude women who are not likely to develop pre-eclampsia from further follow up for the disease (see the low negative Likelihood ratio). On the other hand, formal decision analysis is needed to specify the role of Fn tests as add-ons to clinical information that may usually be available at the point of Fn test ordering decision. For example to answer the question whether it is useful to prescribe preventive drugs to a woman that tested positive.

In conclusion, based on the limited evidence available, the determination of plasma levels of especially cellular Fn seems to be a promising tool to predict pregnant women's risk of pre-eclampsia. Determination of total Fn appears to give a larger variation in results. However, more well-designed and adequately reported studies are necessary to populate ultimate decision-analytic models.

Acknowledgements

This research project was funded by the UK NHS Health Technology Assessment (HTA) Programme (01/64/04).

References

1. Report of the National High Blood Pressure Education Program Working Group on High Blood Pressure in Pregnancy. *Am J Obstet Gynecol.* 2000; 183(1):S1–S22.
2. Cooper GM, Lewis G, Neilson J. Confidential enquiries into maternal deaths, 1997–1999. *Br J Anaesth.* 2002; 89(3):369–72.
3. Montan S, Sjoberg NO, Svenningsen N. Hypertension in Pregnancy – Fetal and Infant Outcome – A Cohort Study. *Clinical and Experimental Hypertension Part B–Hypertens Pregnancy.* 1987; 6(2):337–48.
4. MacGillivray I. Pre-eclampsia. The hypertensive disease of pregnancy.. London: WB Saunders Company Ltd., 1983.
5. Chesley LC. History and epidemiology of pre-eclampsia–eclampsia. *Clin Obstet Gynecol.* 1984; 27(4):801–20.
6. Turnbull AC. Maternal mortality and present trends. In: Sharp F, Symmonds EM, editors. *Hypertension in pregnancy.* Ithaca, New York: Perinatology Press, 1987:135–50.
7. Broughton PF, Crowther C, de Swiet M, Duley L, Judd A, Lilford RJ et al. Where next for prophylaxis against pre-eclampsia? *BJOG.* 1996; 103(7):603–7.
8. Coomarasamy A, Papaioannou S, Gee H, Khan KS. Aspirin for the prevention of pre-eclampsia in women with abnormal uterine artery Doppler: a meta-analysis. *Obstet Gynecol.* 2001; 98(5 Pt 1):861–6.
9. Montan S. Drugs used in hypertensive diseases in pregnancy. *Curr Opin Obstet Gynecol.* 2004; 16(2):111–5.
10. Duley L, Henderson–Smart DJ, Knight M, King JF. Antiplatelet agents for preventing pre-eclampsia and its complications. *Cochrane Database Syst Rev* 2004;(1):CD004659.
11. Roberts JM, Redman CWG. Pre-eclampsia: More than pregnancy-induced hypertension. *Lancet.* 1993; 341(8858):1447–51.
12. Ascarelli MH, Morrison JC. Use of fetal fibronectin in clinical practice. *Obstet Gynecol Surv.* 1997; 52(4 Suppl):S1–12.
13. Dunn PA, Feinberg RF. Oncofetal fibronectin: new insight into the physiology of implantation and labor. *J Obstet Gynecol Neonatal Nurs.* 1996; 25(9):753–7.
14. Koenn ME. Fetal fibronectin. *Clin Lab Sci.* 2002; 15(2):96–8.
15. Honest H, Bachmann LM, Gupta JK, Kleijnen J, Khan KS. Accuracy of cervicovaginal fetal fibronectin test in predicting risk of spontaneous preterm birth: systematic review. *BMJ.* 2002; 325(7359):301.
16. Leitich H, Kaider A. Fetal fibronectin—how useful is it in the prediction of preterm birth? *BJOG.* 2003; 110 Suppl 20:66–70.
17. Conde–Agudelo A, Lede R, Belizan J. Evaluation of methods used in the prediction of hypertensive disorders of pregnancy. *Obstet Gynecol Surv.* 1994; 49(3):210–22.
18. Conde–Agudelo A, Villar J, Lindheimer M. World Health Organization systematic review of screening tests for pre-eclampsia. *Obstet Gynecol.* 2004; 104(6):1367–91.
19. Humpfner A, Hummel S, Heidegger H, Getz B, Deuber HJ, Schulz W. Classes of hypertension in pregnancy – diagnostic possibilities of prediction and differentiation. *Nieren – und Hochdruckkrankheiten* 1994; 23(1):16–32.
20. Magann EF, Martin Jr JN. The laboratory evaluation of hypertensive gravidas. *Obstet Gynecol Surv.* 1995; 50(2):138–45.
21. Myatt L, Miodovnik M. Prediction of pre-eclampsia. *Semin Perinatol.* 1999; 23(1):45–57.
22. O'Brien WF. Predicting pre-eclampsia. *Obstet Gynecol.* 1990; 75(3 Pt 1):445–452.

23. Steinhard J, Klockenbusch W. Pregnancy-induced hypertonia and pre-eclampsia – Risk factors and prediction possibilities. *Gynakologie*. 1999; 32(10):753–60.
24. Visser W, Wallenburg HC. Prediction and prevention of pregnancy-induced hypertensive disorders. *Baillieres Best Pract Res Clin Obstet Gynaecol*. 1999; 13(1):131–56.
25. Young PF, Singh P, O'Brien PMS. Screening for pregnancy-induced hypertension. *Contemp Rev Obstet Gynaecol*. 1994; 6(4):189–94.
26. Zygmunt M, Lang U, Munstedt K. Early prediction of pre-eclampsia: A short review. *Gynakologie*. 2002; 35(7):644–51.
27. Cnossen JS. World Health Organization Systematic Review of Screening Tests for Pre-eclampsia. *Obstet Gynecol*. 2005; 105(5, Part 1):1151–2.
28. Bachmann LM, Coray R, Estermann P, ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *J Am Med Inform Assoc*. 2002; 9(6):653–8.
29. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from MEDLINE: analytical survey. *BMJ*. 2004; 328(7447):1040.
30. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA*. 1994; 271(5):389–91.
31. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999; 282(11):1061–6.
32. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med*. 2004; 140(3):189–202.
33. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003; 3(1):25.
34. Halligan A, Bonnar J, Sheppard B, Darling M, Walshe J. Haemostatic, fibrinolytic and endothelial variables in normal pregnancies and pre-eclampsia. *BJOG*. 1994; 101(6):488–92.
35. Islami D, Shoukir Y, Dupont P, Campana A, Bischof P. Is cellular fibronectin a biological marker for pre-eclampsia? *Eur J Obstet Gynecol Reprod Biol*. 2001; 97(1):40–5.
36. Jones I, Cowley D, Andersen M, Vacca A, Voroteliak V. Fibronectin as a predictor of pre-eclampsia: a pilot study. *Aust N Z J Obstet Gynaecol*. 1996; 36(1):1–3.
37. Ostlund E, Hansson LO, Bremme K. Fibronectin is a marker for organ involvement and may reflect the severity of pre-eclampsia. *Hypertens Pregnancy*. 2001; 20(1):79–87.
38. Paarlberg KM, de Jong CL, van Geijn HP, van Kamp CJ, Heinen AG, Dekker GA. Total plasma fibronectin as a marker of pregnancy-induced hypertensive disorders: a longitudinal study. *Obstet Gynecol*. 1998; 91(3):383–8.
39. Soltan MH, Ismail ZA, Kafafi SM, Abdulla KA, Sammour MB. Values of certain clinical and biochemical tests for prediction of pre-eclampsia. *Ann Saudi Med*. 1996; 16(3):280–4.
40. Sud SS, Gupta I, Dhaliwal LK, Kaur B, Ganguly NK. Serial plasma fibronectin levels in pre-eclamptic and normotensive women. *Int J Gynaecol Obstet*. 1999; 66(2):123–8.
41. Madazli R, Kuseyrioglu B, Uzun H, Uludag S, Ocak V. Prediction of preeclampsia with maternal mid-trimester placental growth factor, activin A, fibronectin and uterine artery Doppler velocimetry. *Int J Gynaecol Obstet*. 2005; 89(3):251–7.
42. Chavarria ME, Lara-Gonzalez L, Gonzalez-Gleason A, Sojo I, Reyes A. Maternal plasma cellular fibronectin concentrations in normal and preeclamptic pregnancies: a longitudinal study for early prediction of pre-eclampsia. *Am J Obstet Gynecol*. 2002; 187(3):595–601.

43. Friedman SA, de Groot CJ, Taylor RN, Roberts JM. Circulating concentrations of fetal fibronectin do not reflect reduced trophoblastic invasion in preeclamptic pregnancies. *Am J Obstet Gynecol.* 1992; 167(2):496–7.
44. Lockwood CJ, Peters JH. Increased plasma levels of ED1+ cellular fibronectin precede the clinical signs of pre-eclampsia. *Am J Obstet Gynecol.* 1990; 162(2):358–62.
45. Taylor RN, Crombleholme WR, Friedman SA, Jones LA, Casal DC, Roberts JM. High plasma cellular fibronectin levels correlate with biochemical and clinical features of pre-eclampsia but cannot be attributed to hypertension alone. *Am J Obstet Gynecol.* 1991; 165(4 Pt 1):895–901.
46. Zhang J, Schisterman EF. Maternal plasma cellular fibronectin for early prediction of pre-eclampsia. *Am J Obstet Gynecol.* 2003; 189(4):1212.
47. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ.* 2003; 326(7379):41–4.
48. Bachmann LM, Ter Riet G, Clark TJ, Gupta JK, Khan KS. Probability analysis for diagnosis of endometrial hyperplasia and cancer in postmenopausal bleeding: an approach for a rational diagnostic workup. *Acta Obstet Gynecol Scand.* 2003; 82(6):564–9.

Accuracy of fibronectin for the prediction of pre-eclampsia





**Galactomannan detection for the diagnosis of
invasive aspergillosis in immunocompromized
patients. A Cochrane Review of Diagnostic Test
Accuracy**

**Mariska M.G. Leeflang, Yvette J. Debets-Ossenkopp,
Caroline E. Visser, Rob J.P.M. Scholten, Lotty Hooft,
Henk A. Bijlmer, Johannes B. Reitsma, Patrick M.M. Bossuyt,
Christina M.J.E. Vandenbroucke-Grauls**

Conducted as a pilot Cochrane Diagnostic Test Accuracy review

Abstract

Background: Invasive aspergillosis (IA) is the most common life-threatening opportunistic invasive mycosis in immunocompromized patients. The main issues in the diagnosis of IA are the following: a test needs to be not too invasive or not too big a burden for the already weakened patient; and a tool is needed to guide therapy. The serum ELISA seems to have potential for both requirements, we therefore wanted to know whether the Platelia ELISA is sufficiently accurate to diagnose IA and to guide antifungal therapy.

Objectives: To obtain summary estimates of the diagnostic accuracy of galactomannan detection in serum for the diagnosis of invasive aspergillosis.

Search strategy: MEDLINE, EMBASE and Web of Science were searched with both Medical Headings and text words for both Aspergillosis and the sandwich ELISA. Furthermore, we tracked references.

Selection criteria: We included studies assessing the diagnostic accuracy of galactomannan detection for the early diagnosis of IA, either prospective or retrospective and either case-control or cohort designs. Patients with neutropenia or patients whose neutrophils are functionally compromised were included. The index test was the Platelia[®] Aspergillus sandwich ELISA, the reference standard was a composite reference standard: EORTC/MSG criteria.

Data collection and analysis: Data collection was done by six reviewers, divided into three pairs of a methodologist and a microbiologist. Data collection and quality assessment was done independently, through a piloted form. Disagreements were solved by discussion.

Results: Seven studies reported results for cut-off value 0.5 ODI. Overall sensitivity was 79% (95% CI 64% to 93%) and overall specificity was 82% (71% to 92%). Twelve studies reported the results for cut-off value of 1.0 ODI, overall sensitivity was 71% (61% to 81%) and overall specificity was 90% (87% to 94%). Seventeen studies reported the results for cut-off value 1.5 ODI, sensitivity was 62% (45% to 79%) and specificity was 95% (92% to 98%).

Authors' conclusions: If we use the test at cut-off value 0.5 in a population of 100 patients with a disease prevalence of 8%, that will mean that 2 patients who have IA, will be missed (sensitivity 79%, 21% false negative rate). And 17 patients will be treated unnecessarily (specificity of 82%, 18% false negative rate). If we use the test at cut-off value 1.5 in the same population, that will mean that 3 IA patients will be missed (sensitivity 62%, 38% false negative rate) and 5 patients will be treated unnecessarily (specificity of 95%, 5% false negative rate).

7.1 Background

7.1.1. Target condition being diagnosed

Invasive aspergillosis (IA) is the most common life-threatening opportunistic invasive mycosis in immunocompromized patients¹. Mortality in patients diagnosed with this condition ranges from 70% to 90% after one year². IA is caused by ubiquitous *Aspergillus* species that invade from (most often) the lungs into the adjacent organs if the immune system is not able to fight the infection. Its incidence is still increasing, mainly because of the increasing number of patients undergoing intensified chemotherapy or receiving prolonged corticosteroid therapy and the increasing number of transplant recipients³⁻⁵.

Establishing the diagnosis of IA in an early stage of infection and subsequent early treatment improves the chances of survival². However, clinical signs and symptoms are non-specific and characteristic lesions on chest radiographs are frequently absent. The only definite reference standard to confirm IA is autopsy, combined with culture from autopsy specimens. As a clinical reference standard, the demonstration of hyphen invasion in tissue specimens obtained by invasive procedures, in combination with a positive culture for *Aspergillus* species from the same specimens, establishes a diagnosis of IA^{6,7}. The problem is that the patient's status often prohibits the use of invasive techniques. Besides that, culturing of the causative agent can result in false negative or false positive results.

In 2001, a committee consisting of the Invasive Fungal Infections Cooperative Group of the European Organization for Research and Treatment of Cancer (EORTC), the Mycoses Study Group (MSG) and the National Institute of Allergy and Infectious Diseases proposed to grade the diagnosis of invasive aspergillosis with three levels of probability of IA⁸: proven, probable and possible IA. Unfortunately, these levels are only useful in research settings, because in clinical practice a large number of patients will be classified as possible IA, which may lead to overexposure to anti-fungal therapy if all possible IA patients are treated⁹.

The main issues in the diagnosis of invasive aspergillosis are the following: a test needs to be sensitive in the early phase of the infection in order to start treatment early, but should not pose a large burden for the already weakened patient. Screening immunocompromized patients for IA weekly or twice a week with such a test may lead to earlier treatment and better outcomes.

Imaging techniques are neither invasive nor too big a burden for most patients. The presence of the so called 'halo sign' or the 'air crescent sign' on radiographs or computed tomography is indicative for IA. These signs are, however, not long-lasting: approximately a week after infection, these signs disappear¹⁰. The costs and the rapid accumulation of radiation associated with CT-scanning prevent its use as a screening tool for IA. Furthermore, imaging techniques only give a clinical diagnosis, not a microbiological diagnosis. Microbiological diagnosis can be achieved

through culturing of the fungus from normally sterile tissues or through histology of those tissues. These techniques, however, are time-consuming and often too invasive for the patient.

An alternative is the use of laboratory tests. These include the detection of antigens (Beta-glucan or galactomannan), measurement of antibodies, or nucleic acid detection techniques. Of these tests, the detection of galactomannan is currently the one that is most often used in practice. Galactomannan is a cell wall component of *Aspergillus* spp. and of *Penicillium* spp.¹¹. It is excreted by the fungus during growth phase and it has been suggested that the level of galactomannan is proportional to the fungal load in tissue and that the level of galactomannan has a prognostic value.

7.1.1 Index test

There are currently two commercially available assays for the detection of galactomannan, the Pastorex[®] latex agglutination test and the Platelia[®] sandwich ELISA test. Of these two, the Pastorex[®] kit is only rarely used nowadays. The ELISA is mostly used for the detection of antigen in serum and in fluid that is obtained via bronchoalveolar lavage (BAL). Other specimens in which the test can also be used are cerebrospinal fluid (CSF) or urine^{6,12}. We focused on the ELISA test in serum, because obtaining serum is less of a burden for the patient than collecting BAL fluid. Results of the ELISA are given as optical density index (ODI), which is the ratio of the optical density of (usually) 1 ng/ml galactomannan versus the optical density of the sample. The cut-off for positivity is recently changed by the manufacturer from 1.5 to 0.5 ODI.

There is substantial variation in the way the galactomannan ELISA is currently used in the clinic. Some clinicians do not use it at all, while others use the galactomannan ELISA as a screening tool, to monitor whether patients at risk develop IA or not. In those cases, serum is tested for IA once or twice every week. Sometimes the galactomannan ELISA is used to test for IA in BAL fluid when IA is already suspected and in those situations, the test is only used in serum when there is no BAL fluid. In most situations, the galactomannan ELISA is used as a triage test: if the ELISA is positive, patients will be referred for further diagnostic testing¹³. The test is also used in the definition of proven, probable or possible IA, or as final decision making tool to start antifungal therapy¹⁴.

7.2 Objectives

Our primary objective was to assess the diagnostic accuracy of galactomannan detection in serum for the diagnosis of invasive aspergillosis in immunocompromized patients, at different cut-off values for test positivity.

7.2.1 Investigation of sources of heterogeneity

We have studied several possible sources of heterogeneity: subgroups of patients, different interpretations of the EORTC/MSG criteria as reference standard, and study design features.

7.3 Methods

7.3.1 Criteria for considering studies for this review

Types of studies

Eligible were studies that assessed the diagnostic accuracy of galactomannan detection by the Platelia[®] sandwich ELISA test, with either prospective or retrospective data collection. The galactomannan ELISA could be assessed alone or in comparison to other tests.

Participants

Studies had to include patients with neutropenia or patients whose neutrophils are functionally compromised. Studies with the following patient groups were included:

- patients with hematological malignancies, receiving hematopoietic stem cell transplants, chemotherapeutics or immunosuppressive drugs;
- solid organ transplant recipients and other patients who were receiving immunosuppressive drugs for a prolonged time;
- patients with cancer who are receiving chemotherapeutics;
- patients with a medical condition compromising the immune system, such as HIV/AIDS and chronic granulomatous disease (CGD), an inherited abnormality of the neutrophils.

Studies that were only regarding HIV/AIDS patients were excluded, because this patient group differs from the other included patient groups in such a way that we regarded them as not being representative.

Index test

The detection of circulating galactomannan was the test under evaluation. We only included studies concerning galactomannan detection in serum. Studies addressing detection in BAL fluid, a number of other body fluids, such as CSF or peritoneal fluid, and tissue were excluded. Two commercially available tests for *Aspergillus* antigen detection are known, a latex agglutination test (Pastorex[®]) and a sandwich ELISA (Platelia[®]). The sandwich ELISA is the most sensitive one and therefore the most widely used. We thus only included studies concerning the ELISA.

Target conditions

The target condition of this review was invasive aspergillosis, also called invasive pulmonary aspergillosis or systemic aspergillosis.

Reference standards

The following clinical reference standards could be used to detect the target condition:

- autopsy;
- the criteria of the EORTC/MSG⁸; or
- the demonstration of hyphal invasion in biopsies, combined with a positive culture for *Aspergillus* species from the same specimens.

The gold standard for this diagnosis is autopsy, combined with a positive culture of *Aspergillus* species from the autopsy specimens, or with histopathological evidence of *Aspergillus*. Because autopsy is rarely reported, we decided to take the criteria of the EORTC/MSG⁸ as the clinical reference standard. These criteria divide the patient population into four categories: patients with proven IA, patients who probably have IA, patients who possibly have IA, and patients without IA. This division is based on host factor criteria, microbiological criteria, and clinical criteria.

7.3.2 Search methods for identification of studies

Electronic searches

The following electronic databases were searched with the search terms mentioned below:

- (1) MEDLINE (through PubMed):
("aspergillosis"[MeSH Terms] OR Aspergillosis[Text Word] OR "aspergillus"[MeSH Terms] OR Aspergillus[Text Word] OR aspergill*) AND ("Nucleic Acid Amplification Techniques"[MeSH] OR Polymerase Chain Reaction[tw] OR PCR[tw] OR nucleic acid amplification[tw] OR immunosorbent assay[tw] OR immunoassay[tw] OR ELISA[tw] OR EIA[tw] OR "immunoassay"[MeSH Terms])
- (2) EMBASE (through OVID):
(exp ASPERGILLOSIS/ or aspergillosis.mp. or aspergillus.mp. or exp ASPERGILLUS/ or aspergill\$.mp.) and (exp Nucleic Acid Amplification/ or nucleic acid amplification.mp. or immunosorbent assay.mp. or exp Enzyme Linked Immunosorbent Assay/ or ELISA.tw. or EIA.tw. or Polymerase Chain Reaction.mp. or exp Polymerase Chain Reaction/ or PCR.tw.)
- (3) Web of Science:
Aspergillosis or aspergillus in title, abstract or subject AND Nucleic Acid Amplification or immunosorbent assay or Enzyme Linked Immunosorbent Assay or ELISA or EIA or Polymerase Chain Reaction or PCR in title, abstract and subject.

Searching other resources

To identify additional published, unpublished and ongoing studies, we

- entered relevant studies identified from the above sources into PubMed and then used the Related Articles feature.
- searched the Science Citation Index to identify articles that cite the relevant articles;
- checked the reference lists of all relevant studies.

In the protocol, we stated that we would also contact authors and industry, but due to time constraints we were not able to do this.

7.3.3 Data collection and analysis

Selection of studies

The first selection, based on title and abstract, was done by one review-author (ML). Articles on animal studies, plant studies, or studies of other fungi than *Aspergillus* were identified at this stage and removed from the search results. Of the remaining articles the full paper was obtained. Three review-authors independently assessed those articles for inclusion (ML, CV, YD). Disagreements were resolved by discussion. Articles on which disagreement could not be resolved were all included.

Data extraction and management

Data were extracted on:

- Author, year of publication and journal;
- Study design;
- Study population;
- Reference standard and performance of the reference standard;
- Performance of the index test;
- Methodological quality;
- Data for two-by-two table.

The data-extraction form was accompanied by a background document that stated how each item on the form should be interpreted. We standardized the form and piloted it on two primary diagnostic studies, including the quality assessment. Data extraction and quality assessment were done by in total six reviewers. Each article was assessed by two review authors, independently. One author had a methodological background and the other a microbiological background. The articles were randomly allocated to a pair of assessors. Disagreements were resolved by discussion.

7.3.4 Assessment of methodological quality

Study quality was assessed using the QUADAS-list, with each item scored as “yes”, “no”, or “unclear”¹⁵. Results are presented in the text, in a graph and in a table. We did not calculate a summary score estimating the overall quality of an article since the interpretation of such summary scores is problematic and potentially misleading^{16,17}.

The items of the QUADAS tool and their interpretation were as follows:

(1) Representative patient spectrum?

We made an inventory of whether patients were inpatients or outpatients, the age groups of the patients and the cause of their increased risk for IA (neutropenia, corticosteroids etc). Furthermore, we assessed how patients were selected for the study. We considered the patient spectrum to be representative when neutropenic patients were consecutively selected in a prospective way. A study that compared results in severely ill patients with the results in relatively healthy patients was not considered representative.

(2) Clear description of selection criteria?

Selection criteria were scored as clearly described if at least the department was stated where the patients were recruited, such as a department of haematology or a department of paediatrics. Such a description can serve as a definition of the patient spectrum.

(3) Reference standard is likely to classify the target condition.

As we only included studies with one of the appropriate reference standards, this item was always fulfilled by all included studies. We did register whether the authors of the primary study used the exact criteria of the EORTC/MSG and (if reported) how they were interpreted.

(4) Time between index and reference test.

The calculation of the diagnostic accuracy of a test is more reliable when the time between the Platelia test and the final diagnosis is not too long. If the galactomannan test is negative on day 1 and the patient is diagnosed as having IA on day 20, this test result will be regarded as a false negative result. The patient's true status on day 1, however, was not known in this case and the false negative result may have been a true negative result at that moment. We judged a time interval of less than 15 days as appropriate. There were however two problems: (1) we expected this item to be reported poorly; and (2) the reference test was in most studies a composite reference while the index test was often used as screening tool to monitor whether patients developed IA. So this item was scored as mean(SD) or median(range) time between the galactomannan test result and the diagnosis, as reported by the authors of the primary study.

(5) Was partial verification prevented?

Because most studies were expected to use a composite reference standard (the EORTC/MSG criteria), we defined partial verification as applying this composite reference standard in less than 90% of the patients. Partial verification would have been a problem in studies where only autopsy is used as reference standard, because it is only done when a patient dies and his or her family gives permission.

(6) Was differential verification prevented?

We defined differential verification as applying a different reference standard in more than 10% of the patients. In this definition, we accepted the EORTC criteria as one reference standard, although we are aware that some patients, those with proven IA, will have been diagnosed by culturing or histology, while other patients, such as the probable ones may have been diagnosed by clinical criteria.

(7) Independent index test and reference test.

The galactomannan ELISA is used as major microbiological criterium in the EORTC criteria. We therefore assessed whether authors explicitly mentioned the exclusion of the ELISA from the EORTC criteria.

(8 and 9) Reporting of index and reference test.

We did not only assess whether the index test and the reference test were clearly described, but we also scored what was reported about the execution of both the index test and the reference standard. These items were scored 'yes' if the authors referred at least to an earlier study, or, in case of the index test, if they stated that the test was done according to the manufacturer's instructions.

(10 and 11) Index test blind for reference test results and vice versa.

These items were scored "yes" if the authors stated explicitly that the assessment of the index test was blinded for the EORTC/MSG results and vice versa.

(12) Was clinical information provided to the researchers?

We did not assess this item, because the EORTC/MSG criteria partially rely on clinical information.

(13) Reporting of uninterpretable or intermediate results?

Intermediate results are optical density indices that are neither positive nor negative. So if authors reported for example that a test result below ODI 1.0 was negative and above 1.5 ODI was positive, and they reported the results of these three categories, then this item was scored as "yes". In our protocol it was also stated that reporting about the in- or exclusion of possible IA patients would also be assessed under this item, but in the QUADAS background document it is stated that this item refers to the results of the index test only and not to the results of the reference standard (as the classification 'possible' is). Furthermore, we have made notice of where the probable IA and possible IA patients were included in the analyses, and we excluded all studies from our meta-analysis that did not report data on test results for these patient groups.

(14) Explanation of withdrawals.

In case there were withdrawals from the study, we scored whether or not those were explained.

(15) Study sponsoring

An extra quality item that we assessed was sponsoring of the study by the manufacturer, because these are known to be important sources of bias in intervention studies^{18,19}

7.3.5 Statistical analysis and data synthesis

Our reference standard was the set of EORTC/MSG criteria that can be used to classify patients to one of four groups: proven IA, probable IA, possible IA and no IA. This resulted in a two-by-four table: positive or negative galactomannan test result in each one of the four reference groups. To calculate test accuracy and to reflect the categories that are used in clinical practice to guide further management, we defined the proven and probable patients as having IA and we defined the possible and no IA patients as not having IA, in order to construct two-by-two tables. Studies reporting insufficient data for the construction of a two-by-two table were excluded from the final analyses.

The data of the two-by-two tables were used to calculate sensitivity and specificity for each study. We present individual study results graphically by plotting the estimates of sensitivity and specificity (and their 95% confidence intervals) in both forest plots and the receiver operating characteristic (ROC) space. We used a bivariate random effects approach for the meta-analysis of the pairs of sensitivity and specificity and for the construction of a summary ROC curve²⁰. This summary ROC curve represents the change in diagnostic accuracy according to changes in cut-off value. The bivariate random effects approach enabled us to calculate summary estimates of sensitivity and specificity, while correctly dealing with the different sources of variation: (1) imprecision by which sensitivity and specificity have been measured within each study; (2) variation beyond chance in sensitivity and specificity between studies; and (3) any correlation that might exist between sensitivity and specificity. Covariates can be incorporated in the bivariate model to examine effect of potential sources of bias and variation across subgroups of studies. Because of the bivariate nature of the model effects of covariates on sensitivity and specificity can be modeled separately.

If more than one threshold was reported, we included the two-by-two-tables for all reported thresholds, to be able to do subgroup analyses per threshold category. For the overall analyses and for the regression analyses in which the threshold was used as covariate, we selected one of those thresholds to incorporate in the meta-analysis. In that case, we chose the threshold of 0.5, if reported, because this is the positivity threshold currently recommended by the manufacturer.

Investigations of heterogeneity

Heterogeneity was investigated in first instance through visual examination of forest plots of sensitivities and specificities and through visual examination of the ROC plot of the raw data.

We addressed the following three sources of heterogeneity: effect of cut-off value, effect of the reference standard and existence of clinical subgroups.

a. Effect of cut-off value

A main source of heterogeneity in diagnostic test accuracy reviews are differences in the applied cut-off value between studies. We expected studies to report three different cut-off values: 1.5 ODI (the value previously prescribed by the manufacturer), 0.5 ODI (the value nowadays prescribed by the manufacturer) and 1.0 ODI (an intermediate value). We therefore first investigated what the influence of these cut-off values was on sensitivity and specificity by comparing these subgroups and by including cut-off value as covariate in the meta-regression model.

Some studies defined a positive test result as one single sample that exceeded the cut-off value, while others defined a test result positive when at least two subsequent samples (taken within a week's time) exceeded the cut-off value. The latter was only reported in studies that used the galactomannan ELISA to monitor whether the patients developed IA. The single sample definition was both used in these screening studies and in studies that only tested for galactomannan when there was suspicion of IA (e.g. fever not responsive to antibacterial medication). The impact of single sample versus subsequent sample was examined within the separate cut-off value categories.

b. Effect of the reference standard

Our reference standard consists of the criteria of the EORTC/MSG, as published by Ascioglu et al. in 2002⁸. Before this publication, however, researchers already used a similar classification of patients into four (or sometimes five or six) groups: proven (also called definite), probable, possible (also called suspected), and no IA. These were published on the internet²¹ or in earlier journal publications²²⁻²⁵. We studied whether the use of these criteria resulted in a different diagnostic test accuracy of the galactomannan ELISA by including reference standard as second covariate in the meta-analyses, in addition to cut-off value.

c. Clinical subgroups

We explored the possible influence of clinical subgroups by stratified analyses and by including additional covariates in the regression analyses. These additional analyses were done within the cut-off value subgroups.

The following variables were used as covariate in the meta-analyses:

- children versus adults;
- distinctive groups of patients (e.g. patients at high risk for IA versus patient with low risk for IA; solid organ transplants versus hematological patients);
- use of antifungal prophylaxis (yes versus no);
- use of antifungal therapy (yes versus no);

Sensitivity analyses

To assess whether methodological quality influenced the results we found, we added each individual quality item as a covariate in the bivariate regression model.

7.4 Results

7.4.1 Results of the search

Our search resulted in 651 hits (506 in August 2005 and 145 via an updated search in April 2007), of which 94 studies were eligible for inclusion, based on title and abstract. After assessment of the full text articles, 52 studies were discarded for various reasons (Figure 7.1) (see Appendix). No extra studies were found through additional searches or reference checking. Thus, this review includes 42 relevant articles²⁶⁻⁶⁷.

Included studies

Table 7.1 lists the characteristics of the 42 included studies, containing a total of 271 patients with proven IA, 356 patients with probable IA, 53 patients that were classified in one group as proven or probable IA, 656 patients with possible IA, 5423 patients with no IA, and 33 patients that were classified in one group as possible or no IA. Five studies were retrospective and 22 were prospective. For 15 studies it was unclear whether they were prospective or retrospective.

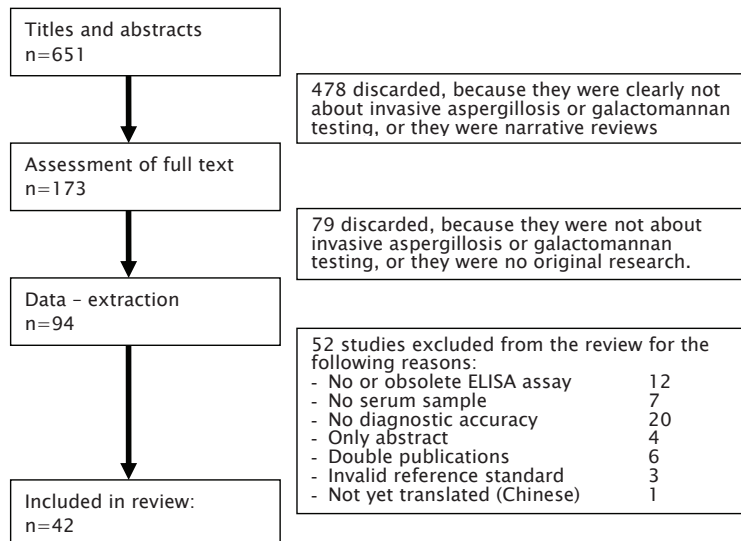


Figure 7.1. Selection of studies

Although we excluded studies that reported in-house assays and obsolete assays, five of these clearly referred to the Stynen 1995 paper, which is the study on which the Platelia[®] galactomannan ELISA was based. We report the results of those studies but have not included them in the analyses.

Ten studies excluded patients who were classified as possible IA. These studies were included in the review, but not in the meta-analyses.

Most studies reported diagnostic accuracy based on the results in individual patients, whereas eight studies reported test results for treatment, neutropenic or disease episodes, without exactly stating how many episodes there were per patient. Because most patients will have only one or two episodes, we did not expect diagnostic test accuracy to change by the inclusion of these studies and therefore we have included those studies in the analyses as well.

Twelve studies directly compared the results of the commercially available galactomannan ELISA with another test (e.g. latex agglutination of galactomannan or PCR), but we did not include these comparisons in our meta-analysis.

The reference standard was formed by the EORTC/MSG criteria that defined the proven, probable, possible or non-IA categories (See Table 7.2). Studies that used criteria that were similar to the EORTC criteria (thus, defining groups of patients with ordinal certainty of IA) were also included.

Excluded studies

Fifty-two articles were excluded (see Characteristics of excluded studies Table in Appendix 7.1). Thirteen assessed another test than the commercially available galactomannan ELISA or an obsolete version, seven studies studied other samples than serum samples, eighteen did not assess diagnostic accuracy at all, four used another reference test, and there were six abstracts of which the results were not yet published in peer reviewed journals. Two studies were double published and one was even published three times. The status of one article in Chinese is uncertain as it has not yet been translated.

7.4.2 Methodological quality of included studies

Figure 7.2 shows the results of the quality appraisal of the 42 included studies. Most studies had included a representative patient spectrum. Five studies reported the results of a case-control study in which they included healthy controls or controls from a different department. These were regarded as not having a representative patient spectrum. Five studies were not clear about how they interpreted the EORTC/MSG criteria or whether they used other criteria as reference standard. The time between the galactomannan ELISA and the actual diagnosis was reported in only five studies, and these reported all an acceptable time gap. Partial and differential verification was no problem. Most studies (n=30) reported explicitly that they did not include the galactomannan ELISA in the EORTC/MSG criteria. Blinding of both

Table 7.1. Characteristics of included studies: country, clinical features, study design, participants, patients per category.

First Author	Country	Clinical features and settings	Study design	Participants	Proven	Probable	Possible	No IA
Versel, 1995	Netherlands	Not reported	Retrospective study	Immunocompromised patients at high risk for IA. No info about age or gender.	6	4	18	33
Sulahan, 1996	France	Inpatients; monitoring clinical course	Prospective; consecutive series of patients.	Patients with hematological malignancies undergoing allogeneic BMT. No info about age or gender.	25	15	8	169
Britagne, 1997	France	Inpatients; monitoring clinical course	Prospective; consecutive series of patients.	Neutropenic children and adults with hematological malignancy or AA, undergoing steroid therapy, or rehospitalized following allogeneic BMT. Age ranged from 8 to 68 years, 66% males.	3	3	9	35
Tabone, 1997	France	In- and outpatients	Prospective; consecutive series of patients.	Immunocompromised children and adolescents from the hematology and oncology pediatric unit. Age ranged from 4 months to 17 years; 57% males.	3	3	2	68
Britagne, 1998	France	Inpatients; monitoring clinical course	Retrospective study	Hematology patients with neutropenia or receiving steroid therapy following allogeneic BMT. No info about age or gender.	6	12	4	19
Macchelli, 1998	Italy	Inpatients; monitoring clinical course	Study design not clear	Patients undergoing allogeneic BMT. No info about age or gender.	1	3	1	17
Ulasakarna, 2000	France	Monitoring clinical course, no further info provided	Consecutive series of patients.	Children and adults undergoing BMT. Age ranged from 6 to 78 years; 67% males.	10	6	2	117
Williamson, 2001	UK	Inpatients; monitoring clinical course	Prospective; consecutive series of patients; blood samples re-analyzed later for ELISA.	Children and adults undergoing BMT or chemotherapy for hematological malignancy with severe neutropenia. Age ranged from 3 months to 56 years. No info about gender.	7	1	9	88
Fertun, 2001	Spain	Not reported	Separate sampling of IA cases and controls; controls were also liver transplants	Adults undergoing liver transplantation and who were not colonized with <i>Aspergillus</i> spp. 'Possibles' excluded. Age ranged from 25 to 66 years. No info about gender.	9 proven or probable			33
Sulahan, 2001	France	Inpatients	Prospective; consecutive series of patients.	Children with hematological malignancies and adults undergoing BMT. 'Possibles' excluded. Age ranged from 6 to 54 years. No info about gender.	32	26	53	744
Baak, 2002	France and Germany	Not reported	Study design not clear.	Children undergoing BMT. Age ranged from 1 month to 9 years; 50% males.	0	1	2	14
Herbrecht, 2002	France	Daily monitoring clinical course, no further info provided	Consecutive series of patients. Episode-based analysis.	Neutropenic children and adults with a persistent fever despite antibiotics, but without any other signs of infection. Age ranged from 4 months to 88 years. No info about gender.	31	67	55	644
Doerrmann, 2002	France	Inpatients; no further info provided	Study design not clear	Adults. No further characteristics reported. Age of cases ranged from 16 to 70 years; no info about gender.	3	9	6	405
Maerrens, 2002	Belgium	Not reported	Prospective; consecutive series of patients.	Adults with hematological disorders who underwent myeloablative ASCT. Exclusion of autologous transplants and patients undergoing nonmyeloablative conditioning. Age ranged from 17 to 58 years; 67% males.	5	8	34	53
Jacque, 2003	Spain	Inpatients	Prospective; consecutive series of patients.	Adults in the hematological department undergoing BMT or chemotherapy. 'Possibles' excluded. Age ranged from 21 to 76 years. No info about gender.	12 proven or probable			88
Moagaes, 2003	Spain	Inpatients; monitoring clinical course	Retrospective study	Severe neutropenic patients in the hematological department. No info about age or gender.	3	1	17	33
Becker, 2003	Netherlands	Inpatients; monitoring clinical course	Prospective; consecutive series of patients.	Adult hematological patients with neutropenia. Age ranged from 18 to 79 years. No info about gender.	2	11	26	48
Kalil, 2003	Tunisia	Inpatients; monitoring clinical course	Prospective; series of patients with same risk profile.	Children and adults that were neutropenic, predominantly allograft patients. Age ranged from 6 to 47 years. No info about gender.	1	4	2	67
Pivik, 2003	France	Inpatients; monitoring clinical course	Prospective; series of patients with same risk profile.	Patients from hematological and intensive care units that were at risk for IFI. No info about age; 62% of the cases were male.	3	31	22	251
Maerrens, 2004	Belgium	Inpatients	Prospective; consecutive series of patients.	Neutropenic adults receiving chemotherapy for AML or MDS, myeloablative allogeneic HSCT, or receiving high-dose steroids for ALL. Age ranged from 16 to 79 years, 61% males.	16	13	21	74
Mari, 2004	USA	Monitoring clinical course, no further info provided	Patients were enrolled prospectively; blood samples were analyzed after storage at -70°C	Children and adults undergoing BMT. Age ranged from 5 to 66 years. No info about gender.	13	11	8	35

Abbreviations: IA = invasive aspergillosis; BMT = Bone Marrow Transplantation; AA = Aplastic anemia; SCT = Stem cell transplantation; ASCT = Autologous SCT; HSCT = Hematopoietic SCT; IFI = Invasive fungal infection; AME = Acute myelogenous leukemia; MDS = Myelodysplastic syndrome; ALE = Acute Lymphoblastic Leukemia; CIVD = Dabé virus for cattle disease.

Table 7.1. (continued) Characteristics of included studies: country, clinical features, study design, participants, patients per category.

First Author	Country	Clinical features and settings	Study design	Participants	Proven	Possible	No IA
Kawazu, 2004	Japan	Weekly screening of inpatients	Prospective, consecutive series of patients. Episode-based analysis.	Adults with hematological disorders that were neutropenic, underwent chemotherapy, had persistent fever despite antibiotics, acute GVHD, or received corticosteroids. Age ranged from 17 to 74 years. 70% males.	9	2	13
Rovira, 2004	Spain	Inpatients were screened twice weekly; outpatients weekly, if possible	Prospective; consecutive series of patients.	Adults undergoing allogeneic HSCT in institution and adult outpatients receiving immunosuppressive therapy. Age ranged from 13 to 60 years. 61% males.	1	5	66
Hassan, 2004	USA	Inpatients	Prospective; consecutive series of patients.	Adults undergoing lung transplantation. No neutropenic patients. No 'Possibles' defined. Age ranged from 21 to 68 years. 47% males.	9	3	58
Buchheid, 2004	Germany	Inpatients; no further info provided	Study design not clear. Episode-based analysis.	Adults with hematological malignancies undergoing chemotherapy by BMT and fulfilled host factor criteria (Aisoglio, 2002). Age ranged from 17 to 81 years. No info about gender.	6	3	93
Quiller, 2004	France	Inpatients; no further info provided	Separate sampling of IA cases and controls.	Children and adults with suspected or confirmed IA and 29 control patients. Age below 77 years. 63% male.	7	19	15
Adams, 2004	France	Inpatients; monitoring clinical course	Prospective; consecutive series of patients.	Adults with hematological malignancies who were likely to be severely neutropenic. Age ranged from 16 to 74 years. No info about gender.	0	2	211
Scomer, 2005	New Zealand	Inpatients; no further info provided.	Prospective; series of patients with same risk profile.	Children and adults undergoing SCT or chemotherapy for hematological malignancy and had fever for >36 hrs. Age of cases ranged from 3 to 79 years. 60% males.	4	1	13
Allan, 2005	Scotland	Not reported	Prospective; series of patients with same risk profile. Episode-based analysis.	Adults undergoing allogeneic or autologous SCT or intensive chemotherapy. Age ranged from 16 to 76 years. No info about gender.	0	1	113
Yoo, 2005	Korea	Inpatients; monitoring of clinical course	Prospective; consecutive series of patients.	Neutropenic adults with fever that did not respond to antibiotic therapy. Possible cases were registered as non-IA. No info about age or gender.	2	12	-
Peñate, 2005	Brazil	Not reported	Prospectively collected serum samples and used after each sample for the subsequent sampling of IA cases and controls.	Children and adults undergoing HSCT and who were suspected of having fungal infections plus 27 control samples. 1 patient younger than 10 years; 1 patient older than 40; rest between 10 and 40 years. 51% males.	1	-	10
White, 2005	UK	Not reported	Study design not clear	Patients considered to be at high risk for IA. No info about age or gender.	1	2	4
Marr, 2005	Canada and USA	Not reported	Three different series of patients with same risk profile analyzed retrospectively for GM	Children and adults with hematological malignancies. 'Possibles' excluded. Age ranged from 5 to 66 years. 54% males.	20	26	269
Fayos, 2005	Spain	Inpatients; monitoring clinical course	Prospective; consecutive series of patients. Episode-based analysis.	Adult hematological cancer patients at high risk (Prestige, 2000). Age ranged from 18 to 70 years. 58% males.	5	3	29
Weisser, 2005	Switzerland	Inpatients; monitoring clinical course	Prospective; consecutive series of patients. Episode-based analysis.	Adults undergoing autologous or allogeneic HSCT or receiving chemotherapy. Age ranged from 16 to 78 years. 51% males.	20 proven or probable	32	109
Basca, 2006	Italy	Inpatients; monitoring clinical course	Prospective; consecutive series of patients.	All adult patients undergoing allogeneic HSCT. Age ranged from 19 to 70 years. No info about gender.	2	0	65
Sankaranarayanan, 2006	Thailand	Inpatients; monitoring of clinical course	Prospective; consecutive series of patients.	Patients receiving chemotherapy or allogeneic HSCT. All patients older than 16. 48% male.	5	12	33 possible or no IA
Florent, 2006	France	Inpatients; monitoring of clinical course	Consecutive series of patients.	All patients with hematological malignancies and >= 15 yrs old and who had samples collected within 1 week from diagnosis. No info about gender.	4	8	39
Hew, 2007	Finland	Inpatients; monitoring clinical course	Prospective; consecutive series of patients. Episode-based analysis.	Positive patients at the hematology/oncology department. Age ranged from 1 to 16 years. 57% male.	1	1	27
Lik, 2007	Taiwan	Inpatients; no further info provided.	Study design not clear	Patients from Intensive Care Units and hematology oncology departments. Patients older than 16 yrs, who were at high risk for developing IA because of prolonged neutropenia. Possibles excluded. Age ranged from 16 to 76 years. 15% males.	5	9	149
Muermans, 2007	Belgium and Netherlands	Inpatients	Retrospective study. Episode-based analysis. Multicenter study.	Patients undergoing HSCT. Age ranged from 4 months to 68 years. 57% male.	19	19	201
Fry, 2007	USA	Inpatients were screened weekly; outpatients when possible	Consecutive series of patients. Retrospective study.	Patients undergoing HSCT. Age ranged from 4 months to 68 years. 57% male.	12 proven or probable	81	28

* Boller et al. Used extra definition that will refer to Ascopic-asep; proven, probable, suspected, possibly, no IA. Abbreviations: IA = invasive aspergillosis; BMT = Bone Marrow Transplantation; AA = Aplastic anemia; SCT = Stem cell transplantation; ASCT = Autologous SCT; HSCT = Hematopoietic SCT; IFI = Invasive fungal infection; AAI = Acute myelogenous leukemia; MDS = Myelodysplastic syndrome; ALL = Acute lymphoblastic leukemia; GVHD = Graft versus host disease.

Table 7.2. Characteristics of included studies: reference standard, index test, cut-off values, sensitivity and specificity.

First Author	Reference Standard to diagnose IA	Description usage index test	Single or subsequent samples required	Cut-offs reported	Sensitivity	Specificity
Vervel, 1995	Unclear. EORTC-like with 4 categories differing from no reference.	Stored samples were tested. At least two samples per pt. One or two consecutive samples were positive; the rest was negative.	Single sample	>1 ng/ml	0.90	0.86
Sulhain, 1996	EORTC-like criteria, no reference. Proven or possible versus no IA.	Sera were sampled weekly the first month and then monthly. Division was made between <10 days before diagnosis and after onset of symptoms. Analytic: any positive = positive, no positive = negative.	Subsequent samples	0.8	0.83	0.81 in BMF patients 0.89 in non-BMF pts
Breitag, 1997	EORTC-like criteria, patients divided into low groups. Reference to Rogers 1990 and (Jung 1991, proven-probable versus no IA.	Serum was prospectively collected at admission and once weekly as long as the risk factors persisted. Difference made between positive (one single positive sample) and secondary positive (two consecutive positive samples) (but not in analysis).	Subsequent samples	>1 ng/ml	1.00	0.89
Tabone, 1997	EORTC-like criteria, no reference.	Samples were collected at admission and thereafter for inpatients weekly and for outpatients at every visit. Positive if two consecutive samples were positive.	Subsequent samples	0.8	1.00	0.89
Breitag, 1998	EORTC-like criteria. Confirmed and probable and suspected versus no IA.	Serum was collected on admission and then once weekly.	Single sample	1 ng/ml	12 true positives out of 22 cases	3 false positives out of 19 controls
Machetti, 1998	EORTC-like criteria. Proven+probable+possible versus no IA.	Serum samples were collected three times a week during the first month and once a week during the second and third month. Positivity was defined as at least two consecutive positive samples. One or less positive was considered negative.	Subsequent samples	Positive if >1.5 and negative if <1.0	0.60	0.82
Uusukanya, 2000	EORTC-like criteria, reference Machetti 1998. Proven+probable+possible versus no IA.	Arthropoemia was monitored weekly. 507 samples from 35 pts were analyzed during 193 neutropenic periods. Positive = one or more positives; negative = all negatives.	Single sample	>1.0	1.00	0.92
Williamson, 2000	EORTC-like criteria, no reference, three groups.	Serum samples were collected and tested twice weekly. One positive is positive; no positive at all is negative.	Single sample	NR	NR	NR
Ferlin, 2001	EORTC-like criteria, reference Deening 1994. Proven and probable versus no IA.	They used stored frozen serum specimens that were routinely obtained. The number of sera per patient varied between 3 and 6. >1 positive is positive; all negative is negative. No possible.	Single sample	>1.0	0.56	0.94
Sulhain, 2001	EORTC-like criteria, based on the retrospective review of clinical charts, pathology, CT scans, and neurological criteria. Proven and probable versus no IA.	Two consecutive positive tests were considered positive. The rest was negative. Suspected patients were excluded.	Subsequent samples	>1.5	0.91	0.94
Bialek, 2002	EORTC criteria, reference EORTC-adviser*. Proven+probable+possible versus no IA.	Screening for aspergillosis (but not reported how often).	Single sample	Positive if >1.5 and negative if <1.0	1.00	0.37
Heibrecht, 2002	EORTC criteria, reference Aiello 2002. Groups analyzed separately.	Serum was collected daily on up to three occasions and then every 7th day; all pts with suggestive clinical signs were tested, not reported what exactly TP and TW was.	Single sample	>1.5	0.65 definite 0.16 probable 0.25 possible	0.95
Doermann, 2002	EORTC-like criteria. Proven+probable+possible versus no IA.	Sera were tested twice weekly in patients at risk. Two consecutive samples is positive, the rest is negative.	Subsequent samples	1.5 ng/ml	0.72	1.00
Marrans, 2002	EORTC criteria, reference Aiello 2002. Proven+probable+possible versus no IA.	Serum samples were collected twice weekly, and more often if patients were proven or probable. Two consecutive was positive. Results were reported back to clinicians once a week...	Subsequent samples	≥1.0	0.94	0.99
Jacqz, 2003	EORTC criteria, reference Aiello 2002. In analyses proven and probable versus no IA.	Twice a week was sampled. Two consecutive positive samples was considered as a positive result. No explanation negative result.	Subsequent samples	≥1.5	0.67	0.97
Mougaert, 2003	EORTC criteria, reference Aiello 2002.	Twice a week was sampled. Two consecutive positive samples was considered as a positive result. No explanation negative result.	Subsequent samples	≥1.5	0.67 (ov) 0.50 (ov+pb) 0.14 (p+pb+pt)	0.67 (ov) 0.50 (ov+pb) 0.14 (p+pb+pt)

* <http://www.aspergillus.man.ac.uk/secure/diagnosis/ident.html>

Table 7.2. (continued) Characteristics of included studies, reference standard, index test, cut-off values, sensitivity and specificity.

First Author	Reference Standard to diagnose IA	Description stage index test	Single or subsequent samples required	Cut-off reported	Sensitivity	Specificity
Becker, 2003	Modified EORTC criteria, they added two extra categories. Reference Ascioglu 2002. Proven and probable and possible versus no IA.	Serum was sampled twice weekly during neutropenia. Two subsequent positive samples were considered positive. Negative(s)?	Subsequent samples	>1.0	0.47	0.93
Kalish, 2003	EORTC criteria. Proven + probable + possible versus no IA.	Sera were monitored weekly on Mondays and Tuesdays. Both days positive = positive. All other results = negative.	Subsequent samples	>1.5	0.71	0.92
Prins, 2003	EORTC criteria, reference Ascioglu 2002. Proven + probable + possible versus no IA.	Two consecutive positive patient samples were necessary to suspect IA.	Subsequent samples	>1.0	0.50	1.00
Maertens, 2004	EORTC criteria, reference Ascioglu 2002. Proven + probable + possible versus no IA.	Serum samples were taken at least twice weekly, until the end of neutropenia, hospitalization or until death. Negative was below 1 and all others were reported as they were (e.g. 1.45). True positive when two consecutive samples were positive.	Single sample	≥0.5 ≥1.0 ≥1.5	0.97 0.93 0.83	0.85 1.00 1.00
Marr, 2004	EORTC criteria, reference Ascioglu 2002.	Blood samples were obtained weekly. Samples were frozen and relabelled randomly. Samples were analysed blinded to both the source of the samples and clinical data. Samples that had absorbance above 0.5 were tested again to verify positive result. At least one sample had to be obtained within 1 week before or after diagnosis.	Subsequent samples	≥0.5 ≥1.0 ≥1.5	0.97 0.79 0.62	0.99 1.00 1.00
Kawachi, 2004	EORTC criteria, reference Ascioglu 2002.	Serum was monitored weekly. Treatment episodes with only one or two measurements were excluded. Positive is either one positive sample or two consecutive positive samples. All the rest is negative.	NR	≥0.5	Total group pr+pb 0.14 pv 0.62 Arbeitsgemeinschaft Pv 0.2; Pb 0.17 No antifungals Pv 0.88; Pb 0.80	Total group poss+no IA 0.74
Rovina, 2004	EORTC criteria, reference Ascioglu 2002. In analyses proven + probable + possible versus no IA.	Serum was monitored twice a week until discharge or death. Outpatients were monitored weekly where possible. Positive was above 1.5 and negative was below 1.0. In between was undetermined. Positive was one or more positive; negative was all negative.	Single sample Subsequent samples	>0.6	1.00	0.55
Husain, 2004	EORTC criteria, reference Ascioglu 2002. In analyses proven and probable versus no IA.	Blood samples were collected twice weekly. Only those tests performed within a week of the diagnosis of IA were considered for the analysis.	Single sample	Positive > 1.5; negative < 1.0	0.75	0.93
Buchheid, 2004	EORTC criteria, reference Ascioglu 2002. Proven and probable versus no IA.	On average every three days, samples were measured. ≥ = 2 consecutive positives was considered positive. Nothing reported about negative results.	Single sample	≥0.5 ≥0.66	0.30 0.30	0.93 0.95
Challier, 2004	EORTC criteria, reference Ascioglu 2002. Children and adults separated; proven and probable versus no IA.	At least two serum samples were obtained from each patient over a maximum of two months before and after suspected infection. One positive sample is positive, the rest is negative?	Subsequent samples	>1.5	0.33	0.99
Adams, 2004	EORTC criteria, reference Ascioglu 2002.	GM antigenemia is monitored weekly. ≥ = 1 positive = positive. All others negative.	Single sample	1 ng/ml	0.75	1.00
Scorier, 2005	EORTC criteria, reference Ascioglu 2002. In analyses proven and probable versus possible and no IA.	If the patient was febrile at least once per day for four days or if there was a high suspicion of IA, tests were attained for galactomannan. Number of samples per patient varied from 2 to 12. At the time point of sample; if any, the patients were classified according to their status at that moment. Negative = all samples negative; positive = at least one sample positive.	Single sample Subsequent	>1.5 ≥0.5<1.0 ≥1.0<1.5 ≥1.5<2.0 ≥2.0 ≥0.5<1.0 ≥1.0<1.5 ≥1.5<2.0 ≥2.0	NR 0.60 0.60 0.60 0.40 0.60 0.40 0.20 0.20	3.8 false positives 0.90 0.90 0.95 1.00 0.90 0.90 0.95 1.00

Table 7.2. (continued) Characteristics of included studies, reference standard, index test, cut-off values, sensitivity and specificity.

First Author	Reference standard to diagnose IA	Description usage index test	Single or subsequent samples required	Cut-offs reported	Sensitivity	Specificity
Ailin, 2005	EORTC criteria, reference Acioglu 2002. No proven IA found. Probable + possible versus no IA.	Twice-weekly screening of serum specimens, both sequential and non-sequential analyzed. Negative = all samples in a certain period were negative.	Single sample	≥0.5	0.00	0.96
				≥1.0	0.00	0.91
				≥1.5	0.25	0.67
			Subsequent samples	≥0.5	0.00	0.99
				≥1.0	0.00	0.99
				≥1.5	0.00	0.90
Yoo, 2005	EORTC criteria, reference Acioglu versus possible and no IA.	Blood samples were usually obtained twice a week until the patient recovered from neutropenia. Two consecutive positive samples was considered as a positive result. No explanation negative result.	Subsequent samples	0.14	0.93	0.65
				0.40	0.86	0.68
				0.41	0.86	0.69
				0.43	0.86	0.71
				0.50	0.86	0.78*
				0.53	0.86	0.72
				0.68	0.79	0.72
Penza, 2005	EORTC criteria, reference Acioglu 2002.	Serum samples collected within a 15-60 days interval in patients with established diagnosis of EI. In patients without any fungal infection, samples were selected at random. The samples selected for this study were stored and frozen for 1 to 5 years.	Single sample	Positive if >1.5 and negative if <1.0	NR	NR
White, 2005	EORTC criteria, no reference.	One positive is positive, no positive is negative, not reported how often samples were taken.	Single sample	>1.5	NR	1.00
Mun, 2005	EORTC criteria, reference Acioglu 2002. Proven and probable versus no IA.	Blood samples collected twice weekly. Patient was positive if ever had positive test result. Negative was having none positive results. Data set restricted to serum samples obtained within 14 days before or after diagnosis. 'Possibles' excluded.	Single sample	>1.5	0.43	0.93
				>1.0	0.48	0.88
				>0.5	0.70	0.70
Fazio, 2005	EORTC criteria, reference Acioglu versus no IA.	Samples were collected twice weekly until the risk for EI had ended, true positive when two consecutive samples were positive. Rest is negative.	Subsequent samples	>1.5	0.88	0.90
Weisser, 2005	EORTC criteria.	Sera were tested twice weekly. Two consecutive positive was considered positive.	Subsequent samples	≥0.5	0.65 (possible) 0.60 (proven or probable)	0.82 (no IA)
Baca, 2006	EORTC criteria, reference Acioglu 2002.	Serum samples were taken twice weekly. ACA positivity was defined as an OD index of galactomannan ≥1.0 in two subsequent sera.	Subsequent samples	≥1.0	1.00	0.93
Sankralay, 2006	EORTC criteria, reference Acioglu 2002.	Blood samples were obtained once or twice weekly until death or discharge.	Subsequent samples	≥0.5	0.94	0.67
				≥0.75	0.94	0.79
				≥1.0	0.88	0.97
				≥1.25	0.88	1.00
				≥1.5	0.77	1.00
Florent, 2006	EORTC criteria, reference Acioglu 2002.	Twice weekly testing for galactomannan.	Single sample Subsequent samples	≥1.5	0.75	0.73
				≥0.5	0.67	0.75
Hew, 2007	EORTC criteria, reference Acioglu 2002.	Sera were tested once a week. Antigen levels were recorded as negative, positive, borderline.	Single sample Subsequent samples	NR	NR	NR
Lai, 2007	EORTC criteria, reference Acioglu 2002.	No information, except about definition of positive test result and cut-off values.	Subsequent samples	≥1.5	0.79	0.94
Maertens, 2007	EORTC criteria, reference Acioglu 2002.	Blood samples were tested twice weekly or once daily.	Single sample	≥1.5	0.76	0.98
				≥1.0	0.82	0.97
				≥0.5	0.97	0.91
				≥0.5	0.92	0.98
Foy, 2007	EORTC/NIHID criteria (no reference)	Biweekly serum samples.	Single sample	>0.5	Children: 0.80 Adults: 0.25	Children: 0.98 Adults: 0.91

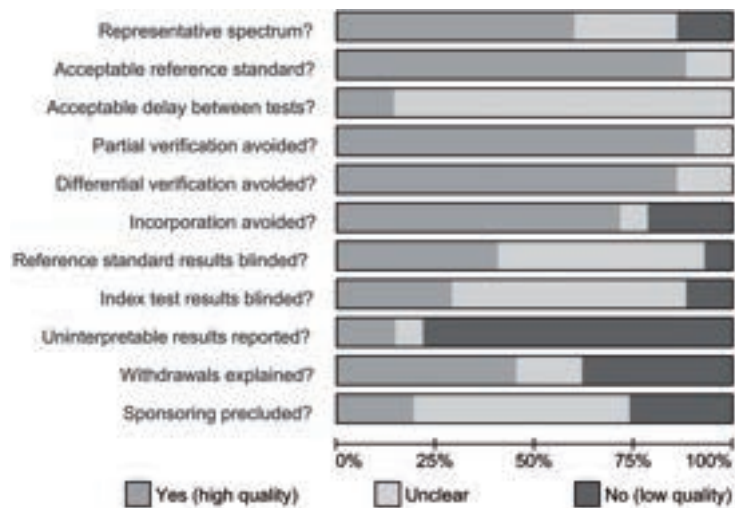


Figure 7.2. Overall quality assessment of all 42 included studies.
Data presented as stacked bars, representing the percentage of studies scoring 'yes', 'unclear' or 'no' on the particular quality item.

the results of the reference standard and the results of the index test was reported variably. Most studies ($n=33$) reported no details at all about any uninterpretable or indeterminate index test results. Explanation of withdrawals was reported 15 studies. Financial support and blinding of both the index and the reference test were poorly reported: 23 studies reported no details.

We did not expect any effects from partial or differential verification, because all studies use more or less the same reference standard in more or less the same way. Factors that may cause bias in our review are: incorporation bias, no representative spectrum, no blinding of both the index and the reference test, no explanation of withdrawals and support by the manufacturer.

7.4.3 Findings

The sensitivity of the 42 studies varied from 0% to 100% and the specificity from 50% to 99% (see Figure 7.3). The wide range of the sensitivity was largely due to chance variation, because of small numbers of patients with the target condition (proven or probable) in the various studies, ranging from 1 to 98 (median 12). For instance, if there is only one patient with proven or probable IA in a study and this patient had a positive test, the sensitivity would be 100%, but if he or she had a negative test result, the sensitivity would be 0%. Small numbers of patients were no issue in the possible or no IA groups (median 95, range 16 to 797).

The median prevalence of IA patients was 12% (range 0.8% to 44%). This prevalence is based on the proportion of proven and probable patients in the studies that in-

cluded consecutive series of patients with a comparable risk to develop IA (in contrast to case-control studies, where the numbers of cases and controls, and thus the prevalence, is determined by the researchers).

Statistical analysis and data synthesis

Nine studies excluded patients with possible IA from their analyses and did not report any data on test performance in this patient group. In theory, the exclusion of patients with “possible IA”, which can be regarded as group of “difficult or atypical” patients, is likely to affect the observed diagnostic accuracy of a test. These studies were therefore excluded from the analyses. We also excluded studies that used an old variant of the galactomannan assay from the meta-analyses.

Thus, our final data set for analysis contained 30 studies. Six of these studies reported on results per disease-episode or treatment-episode, the other reported data per patient. Because there are no reasons to suspect that including the episode-based studies will bias the results, we analyzed all those studies together.

Figure 7.3. Plot of sensitivity versus specificity for all 30 studies, irrespective of cut-off value. The width of the blocks is proportional to the inverse standard error of the specificity in every study and the height of the blocks is proportional to the inverse standard error of the sensitivity.

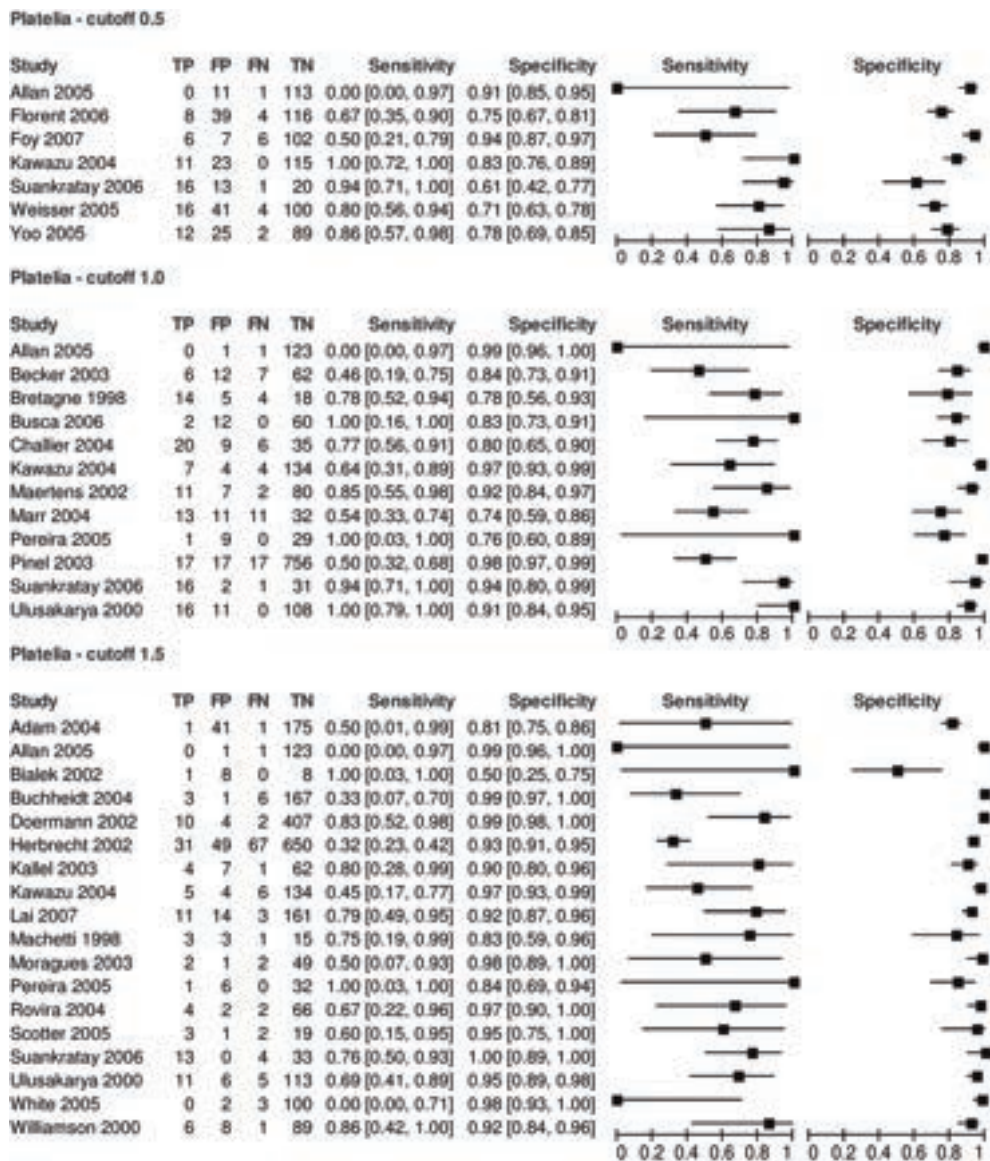


Figure 7.4. Forest plots of sensitivity and specificity.

The squares represent the sensitivity and specificity of one study, the black line its confidence interval. Studies are grouped by reported cut-off value. If a study reported accuracy data for more than one cut-off, its results are included in more than one subgroup.

Of these 30 studies, 20 studies only reported one pair of sensitivity and specificity. These studies only looked at one cut-off value and did not analyse children and adults separately. The remaining studies reported data on more than one cut-off (5 studies), on results when subsequent positive results were needed versus single

Figure 7.5. Summary ROC plots of sensitivity and specificity for different cut-off values (0.5 ODI, 1.0 ODI and 1.5 ODI).
The width of the rectangles is proportional to the number of patients with possible or without IA; the height of the blocks is proportional to the number of patients with IA (proven or probable). The solid line is the summary ROC curve; the thick black spots are the mean values for sensitivity and specificity; the ellipses around the black spots represent the 95% confidence intervals around the summary estimates.

sample results (4 studies), or on separate results for children and adults (3 studies).

Investigations of heterogeneity

a. Effect of cut-off value

Figure 7.4 shows sensitivity and specificity for each study, per cut-off value. Figure 7.5 shows the summary ROC curves and the summary estimates plus confidence ellipses for the different cut-off values separately. Seven studies reported results for a cut-off value of 0.5. The mean sensitivity of those studies was 78% (95% CI 61% to 89%) and mean specificity 81% (95% CI 72% to 88%). Twelve studies reported the results for a cut-off value of 1.0, in which the mean sensitivity was 75% (95% CI 59% to 86%) and mean specificity 91% (95% CI 84% to 95%). Seventeen studies reported the results for a cut-off value of 1.5 in which the mean sensitivity was 64% (95% CI 50% to 77%) and mean specificity 95% (95% CI 91% to 97%). See also Table 7.3. Studies that did not report the cut-off value that they used or used another cut-off value than 0.5, 1.0 or 1.5 ODI were excluded from this analysis.

We also used the cut-off value as a continuous covariate in the regression analysis. For this analysis, the study of Hovi et al. was excluded, because it did not report on cut-off values at all.³⁸ Six of the remaining 29 studies had a cut-off value of 0.5, one study used 0.7, nine studies 1.0 and 13 studies applied 1.5. Although there is a clear trend showing lower sensitivities and higher specificities if a higher cut-off values has been applied, the effects of cut-off value were not statistically significant ($P=0.37$ for sensitivity and $P=0.06$ for specificity). There were no differences between the results of this analysis and the subgroup analysis (Table 7.3). Because the number of studies in the cut-off value subgroups was often too low to assess the effect of the other covariates per cut-off subgroup, we assessed the effects of the other sources of heterogeneity by including them in the regression analyses additional to cut-off value.

Effect of subsequent testing versus single sample testing.

A patient could be defined as test positive in two ways: a single sample above the cut-off value; or two subsequent samples above the cut-off value (Table 7.4).

Table 7.3. Effect cut-off value.

Cut-off	Analysis	Studies (n)	Sensitivity (95% CI)	Specificity (95% CI)
0.5	Subgroup	7	0.78 (0.61 to 0.89)	0.81 (0.72 to 0.88)
	Cut-off as covariate	7	0.78 (0.64 to 0.91)	0.84 (0.75 to 0.93)
1.0	Subgroup	12	0.75 (0.59 to 0.86)	0.91 (0.84 to 0.95)
	Cut-off as covariate	9	0.72 (0.63 to 0.82)	0.89 (0.85 to 0.93)
1.5	Subgroup	17	0.64 (0.50 to 0.77)	0.95 (0.91 to 0.97)
	Cut-off as covariate	12	0.67 (0.51 to 0.83)	0.93 (0.89 to 0.97)

Table 7.4. Effect of definition of test positivity.

Cut-off	Analysis	Studies (n)	Sensitivity (95% CI)	Specificity (95% CI)
0.5	All studies	7	0.78 (0.61 to 0.89)	0.81 (0.72 to 0.88)
	Single Sample	1	0.94 (0.82 to 1.00)	0.61 (0.30 to 0.91)
	Subsequent samples	6	0.74 (0.60 to 0.88)	0.83 (0.77 to 0.90)
	<i>P-value</i>		<i>0.15</i>	<i>0.09</i>
1.0	All studies	12	0.75 (0.59 to 0.86)	0.91 (0.84 to 0.95)
	Single Sample	6	0.83 (0.70 to 0.95)	0.84 (0.74 to 0.94)
	Subsequent samples	6	0.61 (0.51 to 0.81)	0.95 (0.91 to 0.98)
	<i>P-value</i>		<i>0.07</i>	<i>0.02</i>
1.5	All studies	17	0.64 (0.50 to 0.77)	0.95 (0.91 to 0.97)
	Single Sample	10	0.62 (0.44 to 0.81)	0.92 (0.87 to 0.98)
	Subsequent samples	7	0.68 (0.49 to 0.87)	0.97 (0.94 to 0.99)
	<i>P-value</i>		<i>0.65</i>	<i>0.15</i>

b. Effect of reference standard

Most studies (n=17) used the EORTC/MSG criteria to establish a diagnosis of IA and to divide patients into four categories. Six studies did use the EORTC/MSG criteria but divided the patients into two (n=2), three (n=3) or six (n=1) groups. Five studies used criteria that were slightly different from the EORTC/MSG criteria. Four of these divided the patients also in four groups and one study divided the patients into three groups. Using the EORTC criteria was associated with a significantly lower sensitivity ($P=0.03$) without a significant effect on specificity ($P=0.48$), when included as covariate additional to the cut-off value. See Table 7.5. Apparently the criteria affect the proven and probable cases more than the possible and non-IA cases.

c. Clinical subgroups.

It is possible that the accuracy of the following clinical subgroups differs, and therefore act as potential source of heterogeneity:

- children versus adults;
- distinctive groups of patients;
- use of antifungal prophylaxis;
- use of antifungal therapy.

Table 7.5. Effect of reference standard.

Cut-off	Reference Standard	Studies (n)	Sensitivity (95% CI)	Specificity (95% CI)
0.5	EORTC; 4 categories	4	0.69 (0.52 to 0.86)	0.85 (0.75 to 0.95)
	><4 cat.; no EORTC	3	0.86 (0.76 to 0.97)	0.82 (0.70 to 0.94)
1.0	EORTC; 4 categories	6	0.63 (0.51 to 0.75)	0.90 (0.85 to 0.95)
	><4 cat.; no EORTC	4	0.83 (0.73 to 0.93)	0.88 (0.81 to 0.95)
1.5	EORTC; 4 categories	9	0.57 (0.40 to 0.74)	0.93 (0.89 to 0.97)
	><4 cat.; no EORTC	4	0.79 (0.65 to 0.94)	0.92 (0.86 to 0.98)

Table 7.6. Effect of age group.

Cut-off	Studies	Subgroup	Sensitivity (95% CI)	Specificity (95% CI)
0.5	All studies		0.78 (0.64 to 0.91)	0.84 (0.75 to 0.93)
	Foy	Children	0.80 (0.28 to 0.99)	0.98 (0.88 to 1.00)
		Adults	0.29 (0.04 to 0.71)	0.91 (0.81 to 0.96)
	Foy excluded		0.79 (0.64 to 0.93)	0.82 (0.71 to 0.92)
1.0	All studies		0.72 (0.63 to 0.82)	0.89 (0.85 to 0.93)
	Challier	Children	0.92 (0.62 to 1.00)	0.60 (0.36 to 0.81)
		Adults	0.64 (0.35 to 0.87)	0.96 (0.79 to 1.00)
	Challier excluded		0.71 (0.61 to 0.81)	0.90 (0.87 to 0.94)
1.5	All studies		0.67 (0.51 to 0.83)	0.93 (0.89 to 0.97)
	Bialek	Children	1.00 (0.05 to 1.00)	0.50 (0.25 to 0.75)
	Bialek excluded		0.62 (0.45 to 0.79)	0.95 (0.92 to 0.98)

Children versus adults

Within the set of 29 studies there were only three studies that reported data on children. Foy et al.³⁶ and Challier et al.³² reported separate results for both adults and children. Bialek et al.²⁷ reported only results for children. Each of those studies reported a different cut-off value (Table 7.6).

Within the studies that compared directly children with adults, the children showed a higher sensitivity, but the effect on specificity was not straightforward. For this reason, we excluded the (sub)studies that only included children in the further analyses. This resulted in a set of 28 studies.

Effect of distinctive groups of patients

We were not able to investigate the effect of distinctive groups of patients due to the absence of such patient groups in the included studies or this information was not presented in the articles. None of the studies included or reported on patients with solid organ transplantation. Some studies reported the inclusion of high-risk patients, but the definition of high-risk was not always clear or the definition of high-risk matched the inclusion criteria of studies that did not report that they included high-risk studies. Also the type of underlying disease was not always clearly reported.

Therefore, we decided post hoc to analyse the effect of prevalence of IA on the accuracy of the galactomannan test and of the way the patients were selected for the study, as a proxy for disease severity (Table 7.7). High prevalence of IA may reflect a population that is of high risk to develop IA. The effect of prevalence on sensitivity and specificity was not significant when it was in addition to cut-off value as covariate in the regression analysis (sensitivity, $P=0.71$; specificity, $P=0.09$).

Table 7.7. Effect of prevalence.

Cut-off	Prevalence	Studies (n)	Sensitivity (95% CI)	Specificity (95% CI)
0.5	>10%	4	0.78 (0.62 to 0.96)	0.78 (0.66 to 0.90)
	≤10%	2	0.81 (0.63 to 0.99)	0.88 (0.78 to 0.97)
1.0	>10%	6	0.70 (0.56 to 0.84)	0.87 (0.80 to 0.93)
	≤10%	3	0.74 (0.59 to 0.89)	0.93 (0.89 to 0.96)
1.5	>10%	3	0.60 (0.37 to 0.83)	0.92 (0.87 to 0.98)
	≤10%	9	0.65 (0.45 to 0.84)	0.96 (0.94 to 0.98)

Another post-hoc analysis to investigate the effect of distinctive patient groups, was the assessment of the effect of the selection of patients on the accuracy of the galactomannan test. We divided the studies into three groups: (1) studies that did not restrict the patients that would be included in the study and that used the galactomannan ELISA as a screening test in all patients (median prevalence 8.1%, range 0.9 to 12.4%); (2) studies that included only patients who had fever for a certain number of days and whose fever was not responsive to antibiotic treatment (median prevalence 10.9%, range 2.9 to 35.8%); (3) studies that used other selection methods, mostly based on underlying diseases, or that did not report clearly how they selected their patients, or that did use a combination of selection methods (median prevalence 7.4%, range 0.8 to 43.9%). See Figure 7.6. The studies that selected only patients with unresponsive fever reported a significantly lower sensitivity than the other two groups ($P=0.0093$), but with wide confidence intervals. There were no significant differences in specificity (Table 7.8).

Use of antifungal prophylaxis

Fourteen studies used antifungal prophylaxis, 5 studies did not and 9 studies provided no details on the use of prophylaxis. The use of prophylaxis was associated with a significantly lower specificity ($P=0.029$), due to a rise in the proportion of false-positives. Studies using prophylaxis had on average higher sensitivities, but this effect was not significant ($P=0.217$). See Table 7.9.

Table 7.8. Effect of patient selection.

Cut-off	Selection	Studies (n)	Sensitivity (95% CI)	Specificity (95% CI)
0.5	No selection	3	0.73 (0.54 to 0.92)	0.83 (0.69 to 0.96)
	Unresponsive fever	2	0.68 (0.45 to 0.90)	0.88 (0.77 to 0.99)
	Other selection	1	0.89 (0.79 to 0.99)	0.82 (0.67 to 0.96)
1.0	No selection	1	0.61 (0.38 to 0.83)	0.89 (0.87 to 0.98)
	Unresponsive fever	1	0.54 (0.34 to 0.75)	0.93 (0.87 to 0.99)
	Other selection	7	0.82 (0.73 to 0.91)	0.89 (0.82 to 0.95)
1.5	No selection	2	0.47 (0.16 to 0.78)	0.94 (0.87 to 1.00)
	Unresponsive fever	3	0.40 (0.17 to 0.64)	0.96 (0.91 to 1.00)
	Other selection	7	0.72 (0.58 to 0.87)	0.93 (0.89 to 0.97)

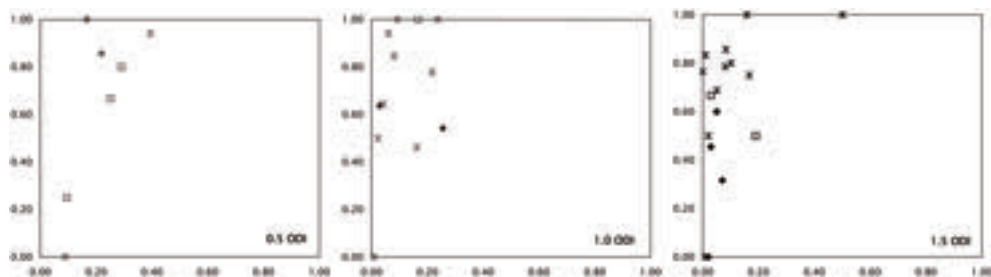


Figure 7.6. Distribution of studies that included only patients with unresponsive fever (black diamonds), studies that included all consecutive patients (grey squares), and patients that used other inclusion criteria (stars) for cut-off values of 0.5 ODI, 1.0 ODI, and 1.5 ODI.

Use of antifungal therapy

Eighteen studies used a therapeutic antifungal intervention (mostly amphotericin B), 2 studies did not and 9 studies were not clear on whether they used therapy or not. Most studies that did use antifungal therapy kept monitoring galactomannan levels during therapy. Use of antifungal therapy increased both sensitivity ($P=0.073$) and specificity ($P=0.047$). See Table 7.10.

Sensitivity analysis

We explored possible sources for bias by adding each individual quality item as covariate in the bivariate regression model, in addition to cut-off value. Quality-items that did have an effect ($P<0.10$) on either sensitivity or specificity were: rep-

Table 7.9. Effect antifungal prophylaxis.

Cut-off	Antifungal prophylaxis	Studies (n)	Sensitivity (95% CI)	Specificity (95% CI)
0.5	no or unclear	2	0.71 (0.51 to 0.92)	0.88 (0.80 to 0.96)
	yes	4	0.82 (0.69 to 0.96)	0.77 (0.65 to 0.89)
1.0	no or unclear	4	0.64 (0.47 to 0.80)	0.93 (0.90 to 0.97)
	yes	5	0.77 (0.64 to 0.89)	0.86 (0.80 to 0.92)
1.5	no or unclear	7	0.55 (0.34 to 0.77)	0.96 (0.94 to 0.99)
	yes	5	0.70 (0.51 to 0.89)	0.92 (0.87 to 0.97)

Table 7.10. Effect antifungal therapy.

Cut-off	Antifungal therapy	Studies (n)	Sensitivity (95% CI)	Specificity (95% CI)
0.5	no or unclear	2	0.64 (0.39 to 0.90)	0.78 (0.67 to 0.90)
	yes	4	0.82 (0.70 to 0.95)	0.89 (0.80 to 0.98)
1.0	no or unclear	1	0.56 (0.34 to 0.77)	0.88 (0.83 to 0.93)
	yes	8	0.76 (0.66 to 0.87)	0.94 (0.90 to 0.98)
1.5	no or unclear	6	0.46 (0.21 to 0.71)	0.93 (0.90 to 0.97)
	yes	6	0.69 (0.52 to 0.86)	0.97 (0.95 to 0.99)

representative patient spectrum (increased sensitivity, $P=0.07$) and blinding of the reference standard (decreased sensitivity, $P=0.06$). Whether or not the results of the galactomannan ELISA were excluded from the EORTC criteria was not significant (sensitivity $P=0.43$; specificity $P=0.14$). Whether or not the study was supported by Platelia[®] was not significant either (sensitivity $P=0.51$; specificity $P=0.89$).

7.5 Discussion

7.5.1 Summary of main results

We included 42 studies in the review, but the results of the meta-analyses are based on the 29 studies that explicitly reported the use of the commercially available galactomannan ELISA, the cut-off value(s), and that included results for all four categories of IA patients: proven, probable, possible, no IA. Quality features that were poorly reported were: the time between the galactomannan ELISA and the actual diagnosis, whether reference and index tests were performed in a blinded fashion, and the source of funding.

The mean sensitivity (children-studies excluded) of the galactomannan ELISA at a cut-off of 0.5 ODI was 79% (95% CI 64% to 93%) and mean specificity 82% (95% CI 71% to 92%). Mean sensitivity and specificity at 1.0 ODI were 71% (95% CI 61% to 81%) and 90% (95% CI 87% to 94%), respectively, and at 1.5 ODI 62% (95% CI 45% to 79%) and 95% (95% CI 92% to 98%), respectively. Especially sensitivity was very heterogeneous. Part of this heterogeneity can be explained by the inclusion of small studies and by the inclusion of studies with low prevalence. See Table 7.11.

Studies that used the EORTC/MSG criteria from 2002 as reference standard had a significantly lower sensitivity. In children, the sensitivity was higher than in adults. Studies that only included patients with fever that was unresponsive to antibacterial therapy, reported a lower sensitivity than other studies. Prevalence of IA had no effect on sensitivity or specificity, but antifungal prophylaxis or therapy did (prophylaxis was associated with a significantly lower specificity; therapy was associated with a significantly higher specificity). Quality items had no significant influence on sensitivity or specificity. See Table 7.12.

Our results compared with other reports

Several reviews have been published in recent years about the (lack of) usefulness of the galactomannan ELISA for the diagnosis of invasive aspergillosis^{14,68-70}. Most of these reviews, however, are based on non-systematic methods. Pfeiffer and colleagues undertook a systematic approach to summarize all available studies until 2005⁶⁸. Although this meta-analysis has methodological limitations (sensitivity and specificity were summarized separately, for example), their results for the different cut-off value subgroups did not differ much from ours⁷¹. Because a change in cut-off value will always lead to an opposite change in sensitivity and specificity

across studies, we studied the effect of other potential factors by including them as covariate additional to the cut-off value. This gives a more realistic estimation of the sensitivity and specificity belonging to a certain group of studies. Pfeiffer et al. also recommended that a higher rather than a lower cut-off value improves diagnostic test accuracy. They only looked, however, at the diagnostic odds ratio (DOR) for this conclusion. Using the DOR to guide clinical decisions regarding the use of a diagnostic test has some serious limitations. It does not take into account the relative importance of false negative or false positive results. A test with a sensitivity of 70% and a specificity of 90% has the same DOR as a test with a sensitivity of 90% and a specificity of 70%, but the clinical consequences of missing a diseased patient (false negative) are not identical to those of given unnecessary treatment to a non diseased patient (false positive).

7.5.2 Strengths and weaknesses of the review

We reviewed the diagnostic accuracy of a commercially available galactomannan ELISA to diagnose invasive aspergillosis (IA) according to the most recent insights and methods for diagnostic meta-analyses. The results can however be biased by the use and implementation of the reference standard, in a way that we have not been able to detect. We only included studies that used the EORTC/MSG criteria or a similar reference standard, but we can imagine that these criteria may still be interpreted subjectively, especially regarding the host factor criteria. Differences in interpretation of the reference standard may have been the reason for the large differences we found in the distribution of patients with proven, probable, possible and non-IA. A relatively large proportion of proven and probable patients may suggest that the reference standard is interpreted in a liberal way, which would then lead to more patients with proven/probable IA that in reality might not have IA. In that case, the estimated sensitivity will be lower than the true sensitivity.

Another factor that we could not control is the time between the index test and the reference standard. Because our reference standard was a composite reference standard, the final diagnosis could have been made at several time points and at different time intervals from the index test. If the time between index test and reference standard is too long, the true disease status of the patient may have been changed by the time the reference standard was assessed.

We defined the proven and probable patients as having IA and we defined the possible and no IA patients as not having IA, in order to construct two-by-two tables. It depends on the association between the galactomannan test results and the true underlying IA status in the probables and in the possibles, whether this would have influenced our results.

7.5.3 Applicability of findings to clinical practice and policy

We reviewed the diagnostic accuracy of only one test, but it would have been worthwhile to investigate the relative value of the galactomannan ELISA in addition to all other tests that can be performed. However, the galactomannan test has the

advantage that it is not an invasive test and hence can be assessed in very ill patients. In some patients, it may therefore be the only available test. In that case, this review gives a valuable overview of the possibilities and weaknesses of the test. Furthermore, the current use of the galactomannan ELISA and its place in the clinic differs from place to place. It would therefore have been very difficult to make comparisons that would have been relevant for a broader public.

In some clinics, the galactomannan test is used in addition to the clinical presentation of the patient and thorax radiographs, as a tool to monitor whether the immunocompromized patient develops IA. If a patient has fever and pulmonary symptoms that do not respond to antimicrobial therapy, he or she will be referred for high-resolution CT (HRCT). If the galactomannan test is positive, the patient will also be referred for HRCT; it is generally believed that the galactomannan test becomes positive before clinical signs of aspergillosis develop. Hence, the use of this test will lead to earlier referral for HRCT, before clear symptoms develop, and to earlier treatment, in case the test is positive. This, in turn, may lead to a higher treatment success rate.

This supposed advantage of the galactomannan test, however, leans on three assumptions: (1) the Platelia test is indeed positive before the patient shows signs and symptoms; (2) the HRCT also shows signs of IA at that moment; and (3) earlier treatment results in a higher success rate. Of the 42 studies that we included in our review, 24 did not report any useful information about point in time at which the galactomannan test was positive. Five studies reported that the test was never positive before either CT, diagnosis or clinical signs. The other studies that reported about time between a positive galactomannan test and other tests or clinical signs, reported time periods varying from around 60 days before to around 50 days after any other evidence (either CT, radiology, clinical signs, fever, diagnosis) for aspergillosis. It was not possible to calculate a mean or median time span, or even a probability of the galactomannan test being earlier positive than other diagnostic evidence. So we could not evaluate the probability that the first two assumptions are true.

7.6 Authors' conclusions

7.6.1 Implications for practice

The value of the galactomannan test will depend on the role that the results of this test will play in clinical decisions about starting therapy for aspergillosis. We can compare the cut-off value of 0.5 ODI with that of 1.5 ODI in a group of 100 potential IA patients with a disease prevalence of 12%. In such a population, 12 patients will have proven or probable IA and 88 will not. If we use the test at a cut-off value 0.5, then we would miss three patients with IA (sensitivity 78%, 22% false negative rate). Although these patients will still be monitored for clinical signs in most clini-

cal situations, the expectation is that IA will be detected later. Seventeen patients will be referred unnecessarily for HRCT (specificity of 81%, 19% false positive rate). If we use the test at a cut-off value of 1.5, then we would miss four patients with IA (sensitivity 64%, 36% false negative rate) and four others will be referred for HRCT unnecessarily (specificity of 95%, 5% false positive rate). Clinicians should decide whether the numbers that follow from the use of the test at 0.5 ODI more or less acceptable than the numbers that follow from the use of the test at 1.5 ODI.

7.6.2 Implications for research

This review showed that, although we do have a good estimate of the test accuracy of the galactomannan ELISA for the diagnosis of IA, we have not enough data to estimate its value in clinical practice. Future studies should report the spectrum of patients in which the test is used unambiguously as well as the time between index test result and actual diagnosis, or between the index test result and results of other tests. It would also be helpful if researchers reported more clearly the individual results of the components of the reference standard.

The diagnostic accuracy of the galactomannan ELISA has been evaluated in several studies. It is time now for studies that evaluate this test as monitoring tool, taking into account the time to diagnosis. It would also be useful to investigate the additional value of the galactomannan ELISA on top of the other tests to diagnose IA.

Table 7.11. Summary of Findings (1)

What is the diagnostic accuracy of the galactomannan ELISA for invasive aspergillosis? And what cut-off should we use?

Patients/population	Immunocompromized patients, mostly heamatology patients
Prior testing	Varied, mostly underlying disease or symptoms (fever, neutropenia)
Settings	Mostly hematology or cancer departments, mostly inpatients
Index test	Galactomannan ELISA, a sandwich ELISA for galactomannan
Importance	Depending on the time-gain the test may give
Reference standard	Clinical and microbiological criteria (gold standard is autopsy, but that is nearly never done)
Studies	Patient series or case-control studies, not using an in-house test and not excluding 'possibles' and reporting a cut-off value (n=29). Each study can be present in more than one subgroup.

Subgroup	Effect (95% CI)	No. of participants (studies)	Prevalence	What do these results mean if the overall median?
Cut-off 0.5	Sensitivity 0.79 (0.61 to 0.93)	901 (7)	Median 9.9% (0.8 to 34%)	With a prevalence of 10%, 10 out of 100 patients will develop IA. Of these, 2 will be missed by the Platelia test (22% of 10), but will be tested again. Of the 90 patients without IA, 17 will be unnecessarily referred for HRCT.
	Specificity 0.82 (0.71 to 0.92)			
In children	Sensitivity 0.92 Specificity 0.60	71 (1)	11%	High rate of false positives in children. Better to use other criteria in children? Or a combination of tests?
Cut-off 1.0	Sensitivity 0.71 (0.61 to 0.81)	1744 (12)	Median 12.4% (0.8 to 44%)	With a prevalence of 12%, 12 out of 100 patients will develop IA. Of these, 3 will be missed by the Platelia test (25% of 12), but will be tested again. Of the 88 patients without IA, only 8 will be unnecessarily referred for HRCT.
	Specificity 0.90 (0.87 to 0.94)			
In children	Sensitivity 0.80 Specificity 0.98	32 (1)	37.5%	In this subpopulation of children no evidence of higher rates of false positives.
Cut-off 1.5	Sensitivity 0.62 (0.45 to 0.79)	2600 (17)	Median 7.4% (0.8 to 34%)	With a prevalence of 7.4%, 7 out of 100 patients will develop IA. Of these, 3 will be missed by the Platelia test (36% of 7), but will be tested again. Of the 93 patients without IA, only 5 will be unnecessarily referred for HRCT.
	Specificity 0.95 (0.92 to 0.98)			
In children	Sensitivity 1.00 Specificity 0.50	17 (1)	6%	High rate of false positives in children. Better to use other criteria in children? Or a combination of tests?

Prevalence over all 28 studies (children-studies excluded): 4501 participants; median 7.7 (IQR 4.6 to 14%)

Table 7.12. Summary of Findings (2)

What factors influence the diagnostic accuracy of galactomannan for invasive aspergillosis?

Patients/population	Immunocompromized patients, mostly heamatology patients
Prior testing	Varied, mostly underlying disease or symptoms (fever, neutropenia)
Settings	Mostly hematology or cancer departments, mostly inpatients
Index test	Galactomannan ELISA, a sandwich ELISA for galactomannan
Importance	Depending on the time-gain the test may give
Reference standard	Clinical and microbiological criteria
Studies	patient series or case-control studies, not using an in-house test and not excluding 'possibles' and reporting a cut-off value (n=29). The analyses were done with cut-off value as first covariate and additional characteristics as second covariate.

Subgroup	Second Covariate	Sensitivity	Specificity	Comments
Cut-off 0.5	None	0.79 (0.64 to 0.93)	0.82 (0.71 to 0.92)	In none of the studied situations, sensitivity or specificity exceeded 90%. Test accuracy improves when patients are being treated for aspergillosis.
	Representative spectrum	0.80 (0.68 to 0.93)	0.81 (0.70 to 0.91)	
	Not representative	0.58 (0.30 to 0.87)	0.85 (0.72 to 0.98)	
	No selection	0.73 (0.54 to 0.92)	0.83 (0.69 to 0.96)	
	Unresponsive fever	0.68 (0.45 to 0.90)	0.88 (0.77 to 0.99)	
	Other selection	0.89 (0.79 to 0.99)	0.82 (0.67 to 0.96)	
	Antifungal prophylaxis	0.82 (0.69 to 0.96)	0.88 (0.80 to 0.96)	
	No prophylaxis	0.71 (0.51 to 0.92)	0.77 (0.65 to 0.89)	
	Antifungal therapy	0.82 (0.70 to 0.95)	0.89 (0.80 to 0.98)	
	No therapy	0.64 (0.39 to 0.90)	0.78 (0.67 to 0.90)	
Cut-off 1.0	EORTC criteria used	0.69 (0.52 to 0.86)	0.85 (0.75 to 0.95)	Sensitivity varies from 54% to 83%. Specificity varies from 86% to 94%. Highest sensitivity/specificity combinations are reached in patients receiving antifungal therapy or prophylaxis, and patients that are preselected on basis of other characteristics than fever.
	Other criteria used	0.86 (0.76 to 0.97)	0.82 (0.70 to 0.95)	
	None	0.71 (0.61 to 0.81)	0.90 (0.87 to 0.94)	
	Representative spectrum	0.79 (0.68 to 0.89)	0.89 (0.84 to 0.94)	
	Not representative	0.56 (0.37 to 0.74)	0.92 (0.87 to 0.97)	
	No selection	0.61 (0.38 to 0.83)	0.89 (0.87 to 0.98)	
	Unresponsive fever	0.54 (0.34 to 0.75)	0.93 (0.87 to 0.99)	
	Other selection	0.82 (0.73 to 0.91)	0.89 (0.82 to 0.95)	
	Antifungal prophylaxis	0.77 (0.64 to 0.89)	0.93 (0.90 to 0.97)	
	No prophylaxis	0.64 (0.47 to 0.80)	0.86 (0.80 to 0.92)	
Cut-off 1.5	Antifungal therapy	0.76 (0.66 to 0.87)	0.94 (0.90 to 0.98)	Although a cut-off value of 1.5 ODI results in the highest specificity, sensitivity may be below 50% in some situations.
	No therapy	0.56 (0.34 to 0.77)	0.88 (0.83 to 0.93)	
	EORTC criteria used	0.63 (0.51 to 0.75)	0.90 (0.85 to 0.95)	
	Other criteria used	0.83 (0.73 to 0.93)	0.88 (0.81 to 0.95)	
	None	0.62 (0.45 to 0.79)	0.95 (0.92 to 0.98)	
	Representative spectrum	0.77 (0.59 to 0.95)	0.94 (0.94 to 0.99)	
	Not representative	0.53 (0.35 to 0.72)	0.96 (0.93 to 0.98)	
	No selection	0.47 (0.16 to 0.78)	0.94 (0.87 to 1.00)	
	Unresponsive fever	0.40 (0.17 to 0.64)	0.96 (0.91 to 1.00)	
	Other selection	0.72 (0.58 to 0.87)	0.93 (0.89 to 0.97)	
	Antifungal prophylaxis	0.55 (0.34 to 0.77)	0.96 (0.94 to 0.99)	
	No prophylaxis	0.70 (0.51 to 0.89)	0.92 (0.87 to 0.97)	
	Antifungal therapy	0.69 (0.52 to 0.86)	0.97 (0.95 to 0.99)	
	No therapy	0.46 (0.21 to 0.71)	0.93 (0.90 to 0.97)	
	EORTC criteria used	0.57 (0.40 to 0.74)	0.93 (0.89 to 0.97)	
	Other criteria used	0.79 (0.65 to 0.94)	0.92 (0.86 to 0.98)	

References

1. Kontoyiannis DP, Bodey GP. Invasive aspergillosis in 2002: an update. *Eur J Clin Microbiol Infect Dis.* 2002; 21(3):161–72.
2. Upton A, Kirby KA, Carpenter P, Boeckh M, Marr KA. Invasive aspergillosis following hematopoietic cell transplantation: outcomes and prognostic factors associated with mortality. *Clin Infect Dis.* 2007; 44(4):531–40.
3. Denning DW, Marinus A, Cohen J, Spence D, Herbrecht R, Pagano L, Kibbler C, Kcrrmery V, Offner F, Cordonnier C, Jehn U, Ellis M, Collette L, Sylvester R. An EORTC multicentre prospective survey of invasive aspergillosis in haematological patients: diagnosis and therapeutic outcome. EORTC Invasive Fungal Infections Cooperative Group. *J Infect* 1998; 37:173–80.
4. Marr KA, Carter RA, Crippa F, Wald A, Corey L. Epidemiology and outcome of mould infections in hematopoietic stem cell transplant recipients. *Clin Infect Dis.* 2002; 34(7) 909–17.
5. Lin MT, Lu HC, Chen WL. Improving efficacy of antifungal therapy by polymerase chain reaction-based strategy among febrile patients with neutropenia and cancer. *Clin Infect Dis.* 2001; 33(10):1621–27.
6. Hope WW, Walsh TJ, Denning DW. Laboratory diagnosis of invasive aspergillosis. *Lancet Infect Dis.* 2005; 5(10):609–22.
7. Singh N. Invasive aspergillosis in organ transplant recipients: new issues in epidemiologic characteristics, diagnosis, and management. *Med Mycol.* 2005; 43 Suppl 1:S267–70.
8. Ascoglu S, Rex JH, de Pauw B, Bennett JE, Bille J, Crokaert F et al. Defining opportunistic invasive fungal infections in immunocompromised patients with cancer and hematopoietic stem cell transplants: an international consensus. *Clin Infect Dis* 2002; 34(1):7–14.
9. Subira M, Martino R, Rovira M, Vazquez L, Serrano D, De La Camara R. Clinical applicability of the new EORTC/MSG classification for invasive pulmonary aspergillosis in patients with hematological malignancies and autopsy-confirmed invasive aspergillosis. *Ann Hematol.* 2003; 82:80–2.
10. Caillot D, Mannone L, Cuisenier B, Couaillier JF. Role of early diagnosis and aggressive surgery in the management of invasive pulmonary aspergillosis in neutropenic patients. *Clin Microbiol Infect.* 2001; 7 Suppl 2:54–61.
11. Latge JP, Kobayashi H, Debeaupuis JP, Diaquin M, Sarfati J, Wieruszkeski JM, Parra E, Bouchara JP, Fournet B. Chemical and immunological characterization of the extracellular galactomannan of *Aspergillus fumigatus*. *Infect Immun.* 1994; 62(12):5424–33.
12. Ascoglu S, de Pauw BE, Donnelly JP, Collette L. Reliability of clinical research on invasive fungal infections: a systematic review of the literature. *Med Mycol.* 2001; 39(1):35–40.
13. Maertens J, Glasmacher A, Selleslag D, Ngai A, Ryan D, Layton M et al. Evaluation of serum sandwich enzyme-linked immunosorbent assay for circulating galactomannan during caspofungin therapy: results from the caspofungin invasive aspergillosis study. *Clin Infect Dis.* 2005; 41(1):e9–14.
14. Segal BH, Walsh TJ. Current approaches to diagnosis and treatment of invasive aspergillosis. *Am J Respir Crit Care Med.* 2006; 173(7):707–17.
15. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol.* 2003; 3:25.
16. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA.* 1999; 282(11):1054–60.
17. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol.* 2005; 5:19.

18. Bero L, Oostvogel F, Bacchetti P, Lee K. Factors associated with findings of published trials of drug–drug comparisons: why some statins appear more efficacious than others. *PLoS Med.* 2007; 4(6):e184.
19. Jørgensen AW, Hilden J, Gøtzsche PC. Cochrane reviews compared with industry supported meta–analyses and other meta–analyses of the same drugs: systematic review. *BMJ.* 2006; 333(7572):782.
20. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology* 2005; 58(10):982–90.
21. The website of the fungal research trust. <http://www.aspergillus.man.ac.uk/secure/diagnosis/defin1.html>. Accessed 1st March 2008.
22. Denning DW, Follansbee SE, Scolaro M, Norris S, Edelstein H, Stevens DA. Pulmonary aspergillosis in the acquired immunodeficiency syndrome. *N Engl J Med.* 1991; 324(10): 654–62.
23. Denning DW, Lee JY, Hostetler JS, Pappas P, Kauffman CA, Dewsnap DH, Galgiani JN, Graybill JR, Sugar AM, Catanzaro A, et al. NIAID Mycoses Study Group Multicenter Trial of Oral Itraconazole Therapy for Invasive Aspergillosis. *Am J Med.* 1994; 97(2):135–44.
24. Rogers TR, Haynes KA, Barnes RA. Value of antigen detection in predicting invasive pulmonary aspergillosis. *Lancet* 1990; 336(8725):1210–3.
25. Machetti M, Feasi M, Mordini N, Van Lint MT, Bacigalupo A, Latge JP et al. Comparison of an enzyme immunoassay and a latex agglutination system for the diagnosis of invasive aspergillosis in bone marrow transplant recipients. *Bone Marrow Transplant.* 1998; 21(9):917–21.
26. Adam O, Auperin A, Wilquin F, Bourhis JH, Gachot B, Chachaty E. Treatment with piperacillin–tazobactam and false–positive *Aspergillus galactomannan* antigen test results for patients with hematological malignancies. *Clin Infect Diseases.* 2004; 38(6):917–20.
27. Allan EK, Jordanides NE, McLintock LA, Copland M, Devaney M, Stewart K, Parker AN, Johnson PR, Holyoake TL, Jones BL. Poor performance of galactomannan and mannan sandwich enzyme–linked immunosorbent assays in the diagnosis of invasive fungal infection. *Br J Haematol.* 2005; 128(4):578–9.
28. Becker MJ, Lugtenburg EJ, Cornelissen JJ, Van Der Schee C, Hoogsteden HC, De Marie S. Galactomannan detection in computerized tomography–based broncho–alveolar lavage fluid and serum in haematological patients at risk for invasive pulmonary aspergillosis. *Br J Haematol.* 2003; 121(3):448–57.
29. Bialek R, Moshous D, Casanova JL, Blanche S, Hennequin C. *Aspergillus* antigen and PCR assays in bone marrow transplanted children. *Eur J Med Res.* 2002; 7(4):177–80.
30. Bretagne S, Marmorat–Khuong A, Kuentz M, Latge JP, Bart–Delabesse E, Cordonnier C. Serum *Aspergillus galactomannan* antigen testing by sandwich ELISA: practical use in neutropenic patients. *J Infect.* 1997; 35(1):7–15.
31. Bretagne S, Costa JM, Bart–Delabesse E, Dhedin N, Rieux C, Cordonnier C. Comparison of serum galactomannan antigen detection and competitive polymerase chain reaction for diagnosing invasive aspergillosis. *Clin Inf Dis.* 1998; 26(6):1407–12.
32. Buchheidt D, Hummel M, Schleiermacher D, Spiess B, Schwerdtfeger R, Cornely OA, Wilhelm S, Reuter S, Kern W, Sudhoff T, Morz H, Hehlmann R. Prospective clinical evaluation of a LightCycler–mediated polymerase chain reaction assay, a nested–PCR assay and a galactomannan enzyme–linked immunosorbent assay for detection of invasive aspergillosis in neutropenic cancer patients and haematological stem cell transplant recipients. *Br J Haematol.* 2004; 125(2):196–202.
33. Busca A, Locatelli F, Barbui A, Limerutti G, Serra R, Libertucci D, Falda M. Usefulness of sequential *Aspergillus galactomannan* antigen detection combined with early radiologic

- evaluation for diagnosis of invasive pulmonary aspergillosis in patients undergoing allogeneic stem cell transplantation. *Transplantation proceedings* 2006; 38(5):1610-3.
34. Challier S, Boyer S, Abachin E, Berche P. Development of a serum-based Taqman real-time PCR assay for diagnosis of invasive aspergillosis. *J Clin Microbiol.* 2004; 42(2):844-6.
 35. Doermann F, Accoceberry I, Weill FX, Agape P, Boiron JM, Marit G, Reiffers J, Couprie B. Evaluation of Aspergillus antigen detection in sera by ELISA for diagnosis of invasive aspergillosis in a hematological unit. *J Mycol Med.* 2002; 12(3):131-5.
 36. Florent M, Katsahian S, Vekhoff A, Levy V, Rio B, Marie JP, Bouvet A, Cornet M. Prospective evaluation of a polymerase chain reaction-ELISA targeted to *Aspergillus fumigatus* and *Aspergillus flavus* for the early diagnosis of invasive aspergillosis in patients with hematological malignancies. *J Inf Dis.* 2006; 193(5):741-7.
 37. Fortun J, Martin-Davila P, Alvarez ME, Sanchez-Sousa A, Quereda C, Navas E, Barcena R, Vicente E, Candelas A, Honrubia A, Nuno J, Pintado V, Moreno S. Aspergillus antigenemia sandwich-enzyme immunoassay test as a serodiagnostic method for invasive aspergillosis in liver transplant recipients. *Transplantation* 2001; 71(1):145-9.
 38. Foy PC, van Burik JA, Weisdorf DJ. Galactomannan antigen enzyme-linked immunosorbent assay for diagnosis of invasive aspergillosis after hematopoietic stem cell transplantation. *Biol Blood Marrow Transplant.* 2007; 13(4):440-3.
 39. Herbrecht R, Letscher-Bru V, Oprea C, Lioure B, Waller J, Campos F, Villard O, Liu KL, Natarajan-Ame S, Lutz P, Dufour P, Bergerat JP, Candolfi E. Aspergillus galactomannan detection in the diagnosis of invasive aspergillosis in cancer patients. *J Clin Oncol.* 2002; 20(7):1898-1906.
 40. Hovi L, Saxen H, Saarinen-Pihkala UM, Vettenranta K, Meri T, Richardson M. Prevention and monitoring of invasive fungal infections in pediatric patients with cancer and hematologic disorders. *Pediatr Blood Cancer.* 2007; 48(1):28-34.
 41. Husain S, Kwak EJ, Obman A, Wagener MM, Kusne S, Stout JE, McCurry KR, Singh N. Prospective assessment of Platelia Aspergillus galactomannan antigen for the diagnosis of invasive aspergillosis in lung transplant recipients. *Am J Transplant.* 2004; 4(5):796-802.
 42. Jarque I, Andreu R, Salavert M, Gomez D, Peman J, Gobernado M, Sanz MA. Value of Aspergillus galactomannan antigen detection in the diagnosis and follow-up of invasive aspergillosis in hematological patients. *Rev Iberoam Micol.* 2003; 20(3):116-8.
 43. Kallek K, Ladeb S, Belhadj S, Jerbi A, Ben Othman T, Abdelkefi A, Ben Abdeladhim A, Chaker E. Evaluation of aspergillary antigenemia for monitoring neutropenic patients. *J Mycol Med.* 2003; 13(4):199-202.
 44. Kawazu M, Kanda Y, Nannya Y, Aoki K, Kurokawa M, Chiba S, Motokura T, Hirai H, Ogawa S. Prospective comparison of the diagnostic potential of real-time PCR, double-sandwich enzyme-linked immunosorbent assay for galactomannan, and a (1 \rightarrow 3)-beta-D-glucan test in weekly screening for invasive aspergillosis in patients with hematological disorders. *J Clin Microbiol.* 2004; 42(6):2733-41.
 45. Lai CC, Hsu HL, Lee LN, Hsueh PR. Assessment of Platelia Aspergillus enzyme immunoassay for the diagnosis of invasive aspergillosis. *J Microbiol Immunol Infect.* 2007; 40(2):148-53.
 46. Machetti M, Feasi M, Mordini N, Van Lint MT, Bacigalupo A, Latge JP, Sarfati J, Viscoli C. Comparison of an enzyme immunoassay and a latex agglutination system for the diagnosis of invasive aspergillosis in bone marrow transplant recipients. *Bone Marrow Transplant.* 1998; 21(9):917-21.
 47. Maertens J, Van Eldere J, Verhaegen J, Verbeken E, Verschakelen J, Boogaerts M. Use of circulating galactomannan screening for early diagnosis of invasive aspergillosis in allogeneic stem cell transplant recipients. *J Inf Dis.* 2002; 186(9):1297-306.
 48. Maertens J, Theunissen K, Verbeken E, Lagrou K, Verhaegen J, Boogaerts M, Eldere JV. Prospective clinical evaluation of lower cut-offs for galactomannan detection in adult neu-

- tropenic cancer patients and haematological stem cell transplant recipients. *Br J Haematol*. 2004; 126(6):852–60.
49. Maertens JA, Klont R, Masson C, Theunissen K, Meersseman W, Lagrou K, Heinen C, Crépin B, Van Eldere J, Tabouret M, Donnelly JP, Verweij PE. Optimization of the cutoff value for the Aspergillus double-sandwich enzyme immunoassay. *Clin Infect Dis*. 2007; 44(10):1329–36.
 50. Marr KA, Balajee SA, McLaughlin L, Tabouret M, Bentsen C, Walsh TJ. Detection of galactomannan antigenemia by enzyme immunoassay for the diagnosis of invasive aspergillosis: variables that affect performance. *J Inf Dis*. 2004; 190(3):641–9.
 51. Marr KA, Laverdiere M, Gugel A, Leisenring W. Antifungal therapy decreases sensitivity of the Aspergillus galactomannan enzyme immunoassay. *Clin Infect Dis*. 2005; 40(12):15.
 52. Moragues MD, Amutio E, Garcia-Ruiz JC, Ponton J. Usefulness of galactomannan detection in the diagnosis and follow-up of hematological patients with invasive aspergillosis. *Rev Iberoam Micol*. 2003; 20(3):103–10.
 53. Pazos C, Ponton J, del Palacio A. Contribution of (1→3)-beta-D-glucan chromogenic assay to diagnosis and therapeutic monitoring of invasive aspergillosis in neutropenic adult patients: a comparison with serial screening for circulating galactomannan. *J Clin Microbiol*. 2005; 43(1):299–305.
 54. Pereira CN, Del Nero G, Lacaz CS, Machado CM. The contribution of galactomannan detection in the diagnosis of invasive aspergillosis in bone marrow transplant recipients. *Mycopathologia* 2005; 159(4):487–93.
 55. Pinel C, Fricker-Hidalgo H, Lebeau B, Garban F, Hamidfar R, Ambroise-Thomas P, Grillot R. Detection of circulating Aspergillus fumigatus galactomannan: value and limits of the Platelia test for diagnosing invasive aspergillosis. *J Clin Microbiol*. 2003; 41(5):2184–6.
 56. Rovira M, Jimenez M, De La Bellacasa JP, Mensa J, Rafel M, Ortega M, Almela M, Martinez C, Fernandez-Aviles F, Martinez JA, Urbano-Ispizua A, Carreras E, Montserrat E. Detection of Aspergillus galactomannan by enzyme immunoabsorbent assay in recipients of allogeneic hematopoietic stem cell transplantation: a prospective study. *Transplantation* 2004; 77(8):1260–4.
 57. Scotter JM, Campbell P, Anderson TP, Murdoch DR, Chambers ST, Patton WN. Comparison of PCR-ELISA and galactomannan detection for the diagnosis of invasive aspergillosis. *Pathology* 2005; 37(3):246–53.
 58. Suankratay C, Kanitcharakul P, Arunyingmongkol K. Galactomannan antigenemia for the diagnosis of invasive aspergillosis in neutropenic patients with hematological disorders. *J Med Assoc Thai*. 2006; 89(11):1851–8.
 59. Sulahian A, Tabouret M, Ribaud P, Sarfati J, Gluckman E, Latge JP, Derouin F. Comparison of an enzyme immunoassay and latex agglutination test for detection of galactomannan in the diagnosis of invasive aspergillosis. *Eur J Clin Microbiol Infect Dis*. 1996; 15(2):139–145.
 60. Sulahian A, Boutboul F, Ribaud P, Leblanc T, Lacroix C, Derouin F. Value of antigen detection using an enzyme immunoassay in the diagnosis and prediction of invasive aspergillosis in two adult and pediatric hematology units during a 4-year prospective study. *Cancer*. 2001;91(2):311–318.
 61. Tabone M-D, Vu-Thien H, Latge J-P, Landman-Parker J, Perez-Castiglioni P, Moissenet D, Leverger G. Value of galactomannan detection by sandwich enzyme-linked immunosorbent assay in the early diagnosis and follow-up of invasive aspergillosis. *Opportunistic Pathogens*. 1997;. 9(1):5–15.
 62. Ulusakarya A, Chachaty E, Vantelon JM, Youssef A, Tancrede C, Pico JL, Bourhis JH, Fenaux P, Munck JN. Surveillance of Aspergillus galactomannan antigenemia for invasive aspergillosis by enzyme-linked immunosorbent assay in neutropenic patients treated for hematological malignancies. *Hematol J*. 2000; 1(2):111–6.

63. Verweij PE, Latge JP, Rijs AJ, Melchers WJ, de Pauw BE, Hoogkamp-Korstanje JA, Meis JF. Comparison of antigen detection and PCR assay using bronchoalveolar lavage fluid for diagnosing invasive pulmonary aspergillosis in patients receiving treatment for hematological malignancies. *J.Clin.Microbiol.* 1995; 33(12):3150-3153.
64. Weisser M, Rausch C, Droll A, Simcock M, Sendi P, Steffen I, Buitrago C, Sonnet S, Gratwohl A, Passweg J, Fluckiger U. Galactomannan does not precede major signs on a pulmonary computerized tomographic scan suggestive of invasive aspergillosis in patients with hematological malignancies. *Clin.Infect.Dis.* 2005; 41(8):1143-1149.
65. White PL, Archer AE, Barnes RA. Comparison of non-culture-based methods for detection of systemic fungal infections, with an emphasis on invasive *Candida* infections. *Journal of clinical microbiology* 2005; 43(5):2181-7.
66. Williamson EC, Oliver DA, Johnson EM, Foot AB, Marks DI, Warnock DW. Aspergillus antigen testing in bone marrow transplant recipients. *Journal of clinical pathology* 2000; 53(5):362-6.
67. Yoo JH, Choi JH, Choi SM, Lee DG, Shin WS, Min WS, Kim CC. Application of nucleic acid sequence-based amplification for diagnosis of and monitoring the clinical course of invasive aspergillosis in patients with hematologic diseases. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2005; 40(3):392-8.
68. Pfeiffer CD, Fine JP, Safdar N. Diagnosis of invasive aspergillosis using a galactomannan assay: a meta-analysis. *Clin Infect Dis.* 2006; 42(10):1417-27.
69. Verdaguer V, Walsh TJ, Hope W, Cortez KJ. Galactomannan antigen detection in the diagnosis of invasive aspergillosis. *Expert Rev Mol Diagn.* 2007; 7(1):21-32.
70. Mennink-Kersten MA, Donnelly JP, Verweij PE. Detection of circulating galactomannan for the diagnosis and management of invasive aspergillosis. *Lancet Infect Dis.* 2004;4(6): 349-57.
71. Leeftang MM, Debets-Ossenkopp YJ, Visser CE, Bossuyt PM. Meta-analysis of diagnostic test accuracy. *Clin Infect Dis.* 2006; 43(9):1220;

Appendix 7.1. Characteristics of excluded studies

Study	Reason for Exclusion
Abdul 2001	No diagnostic test accuracy study
Agape 1999	Double publication
Bart-Delabesse 2005	No diagnostic test accuracy study
Boutboul 2002	No diagnostic test accuracy study
Caillot 2001	No diagnostic test accuracy study
Costa 2002	No diagnostic test accuracy study
Fleck 1999	No galactomannan ELISA
Gari-Toussaint 2001	Invalid reference standard
Giacchino 2006	No diagnostic test accuracy study
Haynes 1990	Obsolete test
Haynes 1994	No galactomannan ELISA
Herrmann 1998	No diagnostic test accuracy study
Hohenthal 2005	No serum (BAL)
Kami 1999	No serum (BAL)
Kami 1999a	No galactomannan ELISA
Kami 2000	No galactomannan ELISA
Kami 2001	Invalid reference standard
Kappe 1996	No galactomannan ELISA
Kwak 2004	No diagnostic test accuracy study
Lim 2004	Abstract
Lombardi 2002	No diagnostic test accuracy study
Maertens 1999	Double publication
Maertens 2001	Double publication
Maertens 2005	No diagnostic test accuracy study
Maesaki 1999	No diagnostic test accuracy study
Marr 2003	Abstract
Mattei 2001	Abstract
Mordini 2001	Abstract
Musher 2004	No serum (BAL)
Pazos 2003	Double publication
Penack 2004	No diagnostic test accuracy study (letter)
Piensi 2001	No galactomannan ELISA
Piensi 2004	No serum (CSF)
Rath 1996	No galactomannan ELISA
Reiss 2000	No diagnostic test accuracy study (review)
Rimek 2002	No galactomannan ELISA
Rogers 1990	Obsolete test
Rohrlich 1996	No diagnostic test accuracy study
Rovira 2003	Double publication
Salonen 2000	No diagnostic test accuracy study
Sanguinetti 2003	No serum (BAL)
Siemann 1998	No diagnostic test accuracy study
Siemann 2001	No diagnostic test accuracy study
Stynen 1995	On this in-house test was the Platelia© based
Ulusakarya 1998	Double publication
Upton 2005	No diagnostic test accuracy study
Verweij 1995a	No serum (BAL)
Viscoli 2002	No serum (CSF)
Viscoli 2004	No diagnostic test accuracy study
Wheat 2007	No diagnostic test accuracy study
White 2006	No galactomannan ELISA
Zedek 2006	Invalid reference standard

Appendix 7.2. References to excluded studies

- Abdul Samad S, Yusoff H, Fadilah SA. The use of an in-house biotin-avidin linked immunosorbent assay to detect *Aspergillus* antigens in sera of immunocompromised patients. *Med J Malaysia*. 2001; 56(1):32-8.
- Agape P, Weill FX, Doermann F, Boiron JM, Pigneux A, Couprie B, Reiffers J, Marit G. Evaluation of serum aspergillus antigen detection using an ELISA method for diagnosis of invasive aspergillosis in patients with hematological malignancies. *Blood* 1999; 94(10):158A.
- Boutboul F, Alberti C, Leblanc T, Sulahian A, Gluckman E, Derouin F, Ribaud P. Invasive aspergillosis in allogeneic stem cell transplant recipients: increasing antigenemia is associated with progressive disease. *Clin Infect Dis*. 2002; 34(7):939-43.
- Caillot D, Mannone L, Cuisenier B, Couaillier JF. Role of early diagnosis and aggressive surgery in the management of invasive pulmonary aspergillosis in neutropenic patients. *Clin Microbiol Infect*. 2001; 7 Suppl 2:54-61.
- Costa C, Costa JM, Desterke C, Botterel F, Cordonnier C, Bretagne S. Real-time PCR coupled with automated DNA extraction and detection of galactomannan antigen in serum by enzyme-linked immunosorbent assay for diagnosis of invasive aspergillosis. *J Clin Microbiol*. 2002; 40(6):2224-7.
- Bart-Delabesse E, Basile M, Al Jijakli A, Souville D, Gay F, Philippe B et al. Detection of *Aspergillus* galactomannan antigenemia to determine biological and clinical implications of beta-lactam treatments. *J Clin Microbiol*. 2005; 43(10):5214-20.
- Fleck E, Rabaud C, Beot S, Chemardin J, Amiel C, May T, Canton P. Invasive pulmonary aspergillosis in AIDS patients. *Med Maladies Infect*. 1999; 29(6):5-15.
- Gari-Toussaint M, Piens MA. Diagnosis of aspergillosis and other invasive filamentous fungal infections in hematology. *Presse Med*. 2001; 30(39-40 Pt 1):1912-7.
- Giacchino M, Chiapello N, Bezzio S, Fagioli F, Saracco P, Alfarano A et al. *Aspergillus* galactomannan enzyme-linked immunosorbent assay cross-reactivity caused by invasive *Geotrichum capitatum*. *J Clin Microbiol*. 2006; 44(9):3432-4.
- Haynes KA, Latge JP, Rogers TR. Detection of *Aspergillus* antigens associated with invasive infection. *J Clin Microbiol*. 1990; 28(9):2040-4.
- Haynes K, Rogers TR. Retrospective evaluation of a latex agglutination test for diagnosis of invasive aspergillosis in immunocompromised patients. *Eur J Clin Microbiol Infect Dis*. 1994; 13(8):670-4.
- Herrmann J, Gugel A, Freidank H, Bertz H, Finke J. *Aspergillus* antigen detection: comparison of a new sandwich ELISA with the latex agglutination test in patients with histologically proven invasive aspergillosis. *Mycoses* 1998; 41 Suppl 1:83-5.
- Hohenthal U, Itala M, Salonen J, Sipila J, Rantakokko-Jalava K, Meurman O, Nikoskelainen J, Vainionpaa R, Kotilainen P. Bronchoalveolar lavage in immunocompromised patients with haematological malignancy--value of new microbiological methods. *Eur J Haematol*. 2005; 74(3):203-11.
- Kami M, Ogawa S, Kanda Y, Tanaka Y, Machida U, Matsumura T, Sakamaki H, Hirai H. Early diagnosis of central nervous system aspergillosis using polymerase chain reaction, latex agglutination test, and enzyme-linked immunosorbent assay. *Br J Haematol*. 1999; 106(2):536-7.
- Kami M, Kanda Y, Ogawa S, Mori S, Tanaka Y, Honda H, Chiba S, Mitani K, Yazaki Y, Hirai H. Frequent false-positive results of *Aspergillus* latex agglutination test - Transient *Aspergillus* antigenemia during neutropenia. *Cancer* 1999; 86(2):274-81.
- Kami M, Tanaka Y, Kanda Y, Ogawa S, Masumoto T, Ohtomo K, Matsumura T, Saito T, Machida U, Kashima T, Hirai H. Computed tomographic scan of the chest, latex agglutination test and plasma (1 -> 3)-beta-D-glucan assay in early diagnosis of invasive pulmonary aspergillosis: A prospective study of 215 patients. *Haematologica* 2000; 85(7):745-52.

Kami M, Fukui T, Ogawa S, Kazuyama Y, Machida U, Tanaka Y, Kanda Y, Kashima T, Yamazaki Y, Hamaki T, Mori S, Akiyama H, Mutou Y, Sakamaki H, Osumi K, Kimura S, Hirai H. Use of real-time PCR on blood samples for diagnosis of invasive aspergillosis. *Clin Infect Dis*. 2001; 33(9):1504-12.

Kappe R, Schulze-Berge A, Sonntag HG. Evaluation of eight antibody tests and one antigen test for the diagnosis of invasive aspergillosis. *Mycoses* 1996; 39(1-2):13-23.

Kwak EJ, Husain S, Obman A, Meinke L, Stout J, Kusne S, Wagener MM, Singh N. Efficacy of galactomannan antigen in the Platelia Aspergillus enzyme immunoassay for diagnosis of invasive aspergillosis in liver transplant recipients. *J Clin Microbiol*. 2004; 42(1):435-8.

Lim ZY, Pagliuca A, Wade J, Ho A, Devereux S, Mufti GJ. Use of the galactomannan ELISA in the detection of invasive aspergillosis: a retrospective review and potential pitfalls in application. *Br J Haematol*. 2004; 125(s1):59-60.

Lombardi G, Farina C, Andreoni S, D'Antonio D, Faggi E, Manso E, Mazzoni A. Multicenter evaluation of an enzyme immunoassay (Platelia Aspergillus) for the detection of Aspergillus antigen in serum. *Mycopathologia* 2002; 155(3):129-33.

Maertens J, Verhaegen J, Demuynck H, Brock P, Verhoef G, Vandenberghe P, Van Eldere J, Verbist L, Boogaerts M. Autopsy-controlled prospective evaluation of serial screening for circulating galactomannan by a sandwich enzyme-linked immunosorbent assay for hematological patients at risk for invasive Aspergillosis. *J Clin Microbiol*. 1999; 37(10):3223-8.

Maertens J, Verhaegen J, Lagrou K, Van Eldere J, Boogaerts M. Screening for circulating galactomannan as a noninvasive diagnostic tool for invasive aspergillosis in prolonged neutropenic patients and stem cell transplantation recipients: a prospective validation. *Blood* 2001; 97(6):1604-10.

Maertens J, Glasmacher A, Selleslag D, Ngai A, Ryan D, Layton M, Taylor A, Sable C, Kartsonis N. Evaluation of serum sandwich enzyme-linked immunosorbent assay for circulating galactomannan during caspofungin therapy: results from the caspofungin invasive aspergillosis study. *Clin Infect Dis*. 2005; 41(1):e9-14.

Maesaki S, Kawamura S, Hashiguchi K, Hossain MA, Sasaki E, Miyazaki Y, Tomono K, Tashiro T, Kohno S. Evaluation of sandwich ELISA galactomannan test in samples of positive LA test and positive aspergillus antibody. *Intern Med*. 1999; 38(12):948-950.

Marr K, Gugel A, Balajee A, Laverdiere M, Laughlin LM, Bentsen C, Leisenring W. A multi-center evaluation of Bio-Rad Platelia Aspergillus GM EIA: Data supporting FDA approval of the test using a low cut-off to define positivity. *Blood* 2003; 102(11):970A.

Mattei D, Mordini N, Gallamini A, Ghirardo R, Ferrua M, Osenda M, Viscoli C. Detection of early stage pulmonary aspergillosis by combined ELISA assay and chest CT scan. *Bone Marrow Transplant*. 2001; 27:S205-S205.

Mordini N, Mattei D, Lo Nigro C, Gallamini A, Ghirardo D, Priotto R, Ferrua MT, Viscoli C. Detection of early stage pulmonary aspergillosis by combined ELISA assay and chest CT scan. *Blood* 2001; 98(11):208A.

Musher B, Fredricks D, Leisenring W, Balajee SA, Smith C, Marr KA. Aspergillus galactomannan enzyme immunoassay and quantitative PCR for diagnosis of invasive aspergillosis with bronchoalveolar lavage fluid. *J.Clin.Microbiol*. 2004; 42(12):5517-22.

Pazos C, del Palacio A. Early diagnosis of invasive aspergillosis in neutropenic patients with bi-weekly serial screening of circulating galactomannan by Platelia Aspergillus. *Rev Iberoam Micol*. 2003; 20(3):99-102.

Penack O, Schwartz S, Thiel E, Wolfgang Blau I. Lack of evidence that false-positive Aspergillus galactomannan antigen test results are due to treatment with piperacillin-tazobactam. *Clin Infect Dis*. 2004; 39(9):1401-2.

Piens MA, Lebeau B, Chapuis F, Thiebaut A, Nicolle MC. Latex Aspergillus antigen detection in invasive aspergillosis: Value in a multicenter prospective study. *J Mycol Med*. 2001; 11(2):61-6.

- Piens M–A, Thiebaut A, Bilger K, De Monbrison F, Michallet M, Picot S. Detection of *Aspergillus* antigen in cerebrospinal fluid in the diagnosis of cerebral aspergillosis. *J Mycol Med.* 2004; 14(1):5–15.
- Rath PM, Oeffelke R, Muller KD, Ansorg R. Non-value of *Aspergillus* antigen detection in bronchoalveolar lavage fluids of patients undergoing bone marrow transplantation. *Mycoses* 1996; 39(9–10):367–70.
- Reiss E, Obayashi T, Orle K, Yoshida M, Zancope–Oliveira RM. Non-culture based diagnostic tests for mycotic infections. *Med Mycol.* 2000; 38 Suppl 1:147–59.
- Rimek D, Kappe R. Invasive aspergillosis: results of an 8-year study. *Mycoses* 2002; 45 Suppl 3:18–21.
- Rogers TR, Haynes KA, Barnes RA. Value of antigen detection in predicting invasive pulmonary aspergillosis. *Lancet* 1990; 336(8725):1210–1213.
- Rohrlich P, Sarfati J, Mariani P, Duval M, Carol A, Saint–Martin C, Bingen E, Latge JP, Vilmer E. Prospective sandwich enzyme–linked immunosorbent assay for serum galactomannan: early predictive value and clinical use in invasive aspergillosis. *Pediatr Infect Dis J.* 1996;15(3):232–7.
- Rovira TM, De la Bellacasa P. *Aspergillus* galactomannan detection in allogenic hematopoietic cell transplantation. *Rev Iberoam Micol.* 2003; 20(3):111–5.
- Salonen J, Lehtonen OP, Terasjarvi MR, Nikoskelainen J. *Aspergillus* antigen in serum, urine and bronchoalveolar lavage specimens of neutropenic patients in relation to clinical outcome. *Scand J Infect Dis.* 2000; 32(5):485–90.
- Sanguinetti M, Posteraro B, Pagano L, Pagliari G, Fianchi L, Mele L, La Sorda M, Franco A, Fadda G. Comparison of real-time PCR, conventional PCR, and galactomannan antigen detection by enzyme–linked immunosorbent assay using bronchoalveolar lavage fluid samples from hematology patients for diagnosis of invasive pulmonary aspergillosis. *J Clin Microbiol.* 2003; 41(8):3922–5.
- Siemann M, Koch–Dorfler M, Gaude M. False–positive results in premature infants with the *Platelia Aspergillus* sandwich enzyme–linked immunosorbent assay. *Mycoses* 1998; 41(9–10):373–7.
- Siemann M, Koch–Dorfler M. The *Platelia Aspergillus* ELISA in diagnosis of invasive pulmonary aspergillosis (IPA). *Mycoses* 2001;44(7–8):266–272.
- Stynen D, Goris A, Sarfati J, Latge JP. A new sensitive sandwich enzyme–linked immunosorbent assay to detect galactofuran in patients with invasive aspergillosis. *J Clin Microbiol.* 1995; 33(2):497–500.
- Ulusakarya A, Chachaty E, Vantelon JM, Bourhis JH, Tancrede C, Hayat M, Pico JL, Munck JN. *Aspergillus* antigenemia surveillance by enzyme–linked immunosorbent assay for aspergillosis in neutropenic patients. *Blood* 1998; 92(10):170A.
- Upton A, Gugel A, Leisenring W, Limaye A, Alexander B, Hayden R et al. Reproducibility of low galactomannan enzyme immunoassay index values tested in multiple laboratories. *J Clin Microbiol.* 2005; 43(9):4796–4800.
- Verweij PE, Stynen D, Rijs AJ, de Pauw BE, Hoogkamp–Korstanje JA, Meis JF. Sandwich enzyme–linked immunosorbent assay compared with *Pastorex latex* agglutination test for diagnosing invasive aspergillosis in immunocompromised patients. *J Clin Microbiol.* 1995; 33(7):1912–4.
- Viscoli C, Machetti M, Gazzola P, De Maria A, Paola D, Van Lint MT, Gualandi F, Truini M, Bacigalupo A. *Aspergillus* galactomannan antigen in the cerebrospinal fluid of bone marrow transplant recipients with probable cerebral aspergillosis. *J Clin Microbiol.* 2002; 40(4):1496–9.
- Viscoli C, Machetti M, Cappellano P, Bucci B, Bruzzi P, Van Lint MT, Bacigalupo A. False–positive galactomannan *Platelia Aspergillus* test results for patients receiving piperacillin–tazobactam. *Clin Infect Dis.* 2004; 38(6):913–6.

Galactomannan detection for the diagnosis of invasive aspergillosis

Wheat LJ, Hackett E, Durkin M, Connolly P, Petraitiene R, Walsh TJ, Knox K, Hage C. Histoplasmosis-associated cross-reactivity in the BioRad Platelia Aspergillus enzyme immunoassay. *Clin Vaccine Immunol.* 2007; 14(5):638-40.

White PL, Linton CJ, Perry MD, Johnson EM, Barnes RA. The evolution and evaluation of a whole blood polymerase chain reaction assay for the detection of invasive aspergillosis in hematology patients in a routine clinical setting. *Clin Infect Dis.* 2006; 42(4):479-486.

Zedek DC, Miller MB. Use of galactomannan enzyme immunoassay for diagnosis of invasive aspergillosis in a tertiary-care center over a 12-month period. *J Clin Microbiol.* 2006; 44(4):1601.





8.1 Summary

Good evidence of the accuracy of diagnostic tests is required to make rational decisions about the provision, selection and application of tests, and to guide the interpretation of test results. Systematic reviews and meta-analyses of test accuracy studies may be the preferred source of such evidence, but building reviews and summarizing study results can be methodologically challenging. The objective of the research reported in this thesis was to provide empirical evidence to guide the development of systematic reviews of diagnostic test accuracy. We specifically addressed the search process, the incorporation of study quality, and the analysis of the data.

Chapter 1 explained the challenges that systematic reviews of diagnostic test accuracy pose and provided an overview of the most recent developments in the methodology for conducting systematic reviews and meta-analyses of diagnostic test accuracy studies. The methods discussed in this Chapter are a reflection of the review methods that will be advocated by The Cochrane Collaboration. The Cochrane Collaboration is the largest international organisation preparing, maintaining and promoting systematic reviews and from October 2008 their Cochrane Database of Systematic Reviews will also include systematic reviews of diagnostic test accuracy.

Diagnostic test accuracy reviews aim to identify and evaluate all available evidence about a specific index test or a comparison of tests. If the yield of the initial search of the literature is too large, a diagnostic search filter can be helpful to reduce this number. The aim of the study reported in **Chapter 2** was to assess the fraction of relevant studies that did not pass methodological search filters for diagnostic test accuracy studies. We also determined to what extent the diagnostic search filters decrease the number of studies that need to be screened to find one relevant article. The use of search filters for diagnostic studies led to an inevitable loss of relevant articles, varying from an average of 2% of the total number of relevant primary articles used in this study to 42%. The major reasons for this loss of articles are the poor indexing of diagnostic studies in MEDLINE and the wide range of possible designs for diagnostic accuracy studies. Search filters are also not guaranteed to reduce the number of studies, so their impact on search efficiency will be small. We feel therefore that the use of diagnostic search filters in the development of a systematic review should be discouraged.

The objective of the research in **Chapter 3** was to examine to what extent different strategies of defining and incorporating quality of included studies affect the results of meta-analyses of diagnostic test accuracy. We re-analyzed the data from 30 systematic reviews by applying three strategies that varied both in the definition of quality and in statistical approach: (1) restricting the analysis to high-quality subsets; (2) multivariable adjustment for a predefined set of quality items; and (3) multivariable adjustment for significant quality items. We found no evidence for

our hypothesis that adjustment for quality in the meta-analysis will lead to less optimistic summary diagnostic accuracy estimates with less variability in results among better-quality studies. The effect of each strategy varied much from one review to another, but also within a single review.

Chapter 4 addressed a possible source of bias when evaluating a test that produces a continuous result: the post-hoc determination of an optimal cut-off value. Optimal cut-off values for continuous test results are often derived in a data-driven way. As this may lead to overoptimistic measures of diagnostic accuracy, we determined the magnitude of bias in sensitivity and specificity associated with data-driven selection of cut-off values in simulated data sets. Three alternative approaches (assuming a specific distribution, leave-one-out, smoothed ROC) were examined for their ability to reduce this bias. The magnitude of bias caused by data-driven optimization of cut-off values was inversely related to sample size. The distribution of the test results had little impact on the amount of bias if sample size was held constant. More robust methods of optimizing cut-off values were less prone to bias, but the performance deteriorated if the underlying assumptions were not met.

Chapter 5 highlighted a possible source of heterogeneity between studies: differences in the prevalence of the target condition. Although it is sometimes claimed that sensitivity and specificity do not depend on disease prevalence, we provide a number of real life examples in which accuracy varied with prevalence. Changes in prevalence and accompanying changes in sensitivity and specificity may be caused by clinical or artefactual variability between studies. Clinical variability refers to differences in the clinical situation. For example, a patient population with a higher disease prevalence may include more severely diseased patients, in which the test performs better. Artefactual variability refers to effects on prevalence and accuracy associated with study design, for example the verification of index test results by a reference standard. Sensitivity and specificity are not fixed test characteristics, but test properties that describe the behaviour of the test in a particular situation. As the setting, filter, or patient group changes, prevalence and accuracy may change. For this reason, variation in disease prevalence and test accuracy between studies can act as a flag for clinicians to detect important differences in study population or study design, affecting accuracy.

In **Chapter 6** we systematically reviewed the accuracy of fibronectin tests for the prediction of pre-eclampsia, one of the most important causes of maternal and fetal mortality and morbidity worldwide. Only five studies reported sufficient data to calculate accuracy estimates, such as sensitivity and specificity. At a sensitivity of at least 50%, specificities ranged between 72 and 96% for cellular fibronectin. For total fibronectin, these numbers were 42 to 94%. Due to the small number of studies and the clinical heterogeneity between studies, we refrained from doing a meta-analysis.

Chapter 7 contained a systematic review of the diagnostic accuracy of galactomannan detection in serum for the diagnosis of invasive aspergillosis (IA) in immunocompromized patients. Twenty-nine studies were included in the meta-analyses. We translated the results of the meta-analyses results to a clinical example. If we use the test at cut-off value 0.5 in a population of 100 patients with a disease prevalence of 8%, that will mean that 2 patients who have IA, will be missed (sensitivity 79%, 21% false negatives) and that 17 patients will be treated unnecessarily (specificity of 82%, 18% false positives). If we use the test at cut-off value 1.5 in the same population, 3 IA patients will be missed (sensitivity 62%, 38% false negatives) and 5 patients will be treated unnecessarily (specificity of 95%, 5% false positives). To improve our understanding of the consequences of false positive and false negative test results in patients, we need more information about (1) the timing of a positive galactomannan test in the course of disease; (2) the timing of positive results in additional tests (for example, clinical signs or CT); and (3) whether earlier treatment improves survival in these patients.

8.2 General discussion

Systematic reviews of diagnostic test accuracy studies are more complicated than systematic reviews of randomized trials. This starts already with question formulation, where the actual or anticipated role of the test in clinical practice and specifications of the patient spectrum are important items to include. In the work reported in this thesis, we specifically addressed the next three steps of a systematic review of diagnostic test accuracy studies: the search process, the incorporation of study quality, and the analysis of the data.

Identification of diagnostic test accuracy studies is complicated by the poor indexing of diagnostic studies in bibliographic databases and the wide range of possible designs for diagnostic accuracy studies. When The Cochrane Collaboration started with its Database of Systematic Reviews and with the developments of systematic reviews of interventions, the same identification problems were encountered for intervention studies. Since then, much effort has gone into the implementation of a clear, unequivocal indexing term for these studies (every randomized controlled trial is now labelled with publication type “randomized controlled trial”) and the development of a register of randomized controlled trials and clinical trials (CENTRAL). One could question whether the same efforts should go into the indexing and registering of diagnostic test accuracy studies.

There is so much variation in diagnostic test accuracy studies that labelling may turn out to be complicated. For example, data on diagnostic test accuracy can be hidden in studies that did not have test accuracy estimation as their primary objective. At this moment, there is no evidence that missing one or two studies will lead to other conclusions. Furthermore, as we saw in Chapter 1, the studies that were not

retrieved by searches were most often older studies, in which outdated diagnostic devices may have been used anyway. In many instances, searching with terms for index test(s) and target condition will suffice.

It may be more efficient to put efforts into promoting informative reporting in individual studies and better implementation of the STARD statement. Complete and transparent reporting of diagnostic test accuracy studies may improve their visibility and thus the retrieval of these studies. Better reporting of study design features is also needed for a better understanding of the relation between methodological quality and biased results. Although there is evidence that individual quality items produce biased results, the intertwined effects of quality items cannot be predicted. Furthermore, the importance of different quality items will vary from one research project to another. Research in these directions is also hampered by poor reporting of study characteristics.

Better reporting of study characteristics may also improve explorations of sources of heterogeneity between studies other than methodological quality. Examples are differences in patient spectrum, in setting or in referral pattern of patients, and in the test under evaluation. Heterogeneity is more a rule than an exception in diagnostic test accuracy reviews, which make random effects models the recommended method for meta-analysis of such data. Heterogeneity can be investigated by including study characteristics as covariates in models for meta-analysis, but drawing conclusions based on these investigations only makes sense if enough information is provided by the included studies.

Troublesome identification of studies and poor reporting of methodological quality and study characteristics complicate the interpretation of the results of meta-analyses of diagnostic test accuracy. Summary estimates of sensitivity and specificity alone provide insufficient information. First, they are no fixed properties of a test. Accuracy measures may change from setting to setting and from population to population. However, to what extent they differ can be difficult to assess, because these changes may be confounded by other, often poorly reported study characteristics and flaws in the methodological design. Second, a test is never used on its own. Diagnostic tests are used to reduce uncertainty about a patient's health status. The results of previous tests may influence the extent to which other tests reduce remaining uncertainty. As long as individual studies do not take combinations of tests into account, reviewing comparative diagnostic questions (add-on, triage, replacement) has to be limited to indirect comparisons between tests, investigated in different patient populations and against different reference standards. Again, the comparative accuracy of these tests may be confounded by other study characteristics. Third, to be able to judge the clinical usefulness of a test or test combination, information should be provided about the consequences of false positive and false negative test results. In case of a false positive test result, the following questions will be important. Will these patients be referred for (invasive) further testing, will they receive a relatively cheap and harmless drug therapy, or will they be referred

for surgery? How many patients on an annual basis will have a false positive test result? In case of false negative test results, we should ask whether this is a severe condition, or not, and will patients be sent home and never seen again, or will they be followed up? Again the question arises how many patients will be involved. When there is an inconsistency or inconclusiveness in the answers to these questions, the studies that are included in the review may provide information about what is usually done in clinical practice, and how the supposed role of the test(s) under evaluation can be applied in those situations.

In conclusion, the development of methods for identification of studies, for the assessment of methodological quality, for meta-analysis, and for the investigation of statistical and clinical heterogeneity will profit from better reporting of design and characteristics of individual studies. Although quality of reporting is not the same as the methodological quality of a study, better reporting of study characteristics will improve the interpretation of study results and thus the overall quality of a study.

The development and conduct of systematic reviews of diagnostic test accuracy is complicated and the methodology is still in progress. Translating the results presented in those reviews into policy and practice may be even more challenging. It requires knowledge of diagnostic research methodology but also of the clinical context in which the tests are used. When reporting of original research improves, we would like to urge authors of diagnostic test accuracy reviews to guide their readers in understanding the implications of their results. Only then can they rest assured that these results will find their way into clinical practice and improve patient care.





Samenvatting en discussie

Samenvatting

Om rationele beslissingen te kunnen nemen over het aanbieden, de selectie en de toepassing van diagnostische tests, is een valide wetenschappelijke onderbouwing van de diagnostische accuratesse van deze tests nodig. De accuratesse van een diagnostische test is het vermogen van die test om personen met en zonder een bepaalde aandoening van elkaar te onderscheiden. De diagnostische accuratesse wordt veelal uitgedrukt in termen als sensitiviteit (het percentage mensen met de aandoening die ook een positieve testuitslag hebben) en specificiteit (het percentage mensen zonder de aandoening die inderdaad een negatieve testuitslag hebben).

Een valide wetenschappelijke onderbouwing van de diagnostische accuratesse van tests of combinaties van tests kan verkregen worden door op systematische wijze de resultaten van eerder gepubliceerde (en niet gepubliceerde) onderzoeken bijeen te brengen en samen te vatten in een overzichtsartikel. Deze systematische reviews van diagnostische accuratesse studies mogen dan de voorkeur genieten voor de onderbouwing van beslissingen boven individuele studies, het schrijven ervan en het samenvatten van de studieresultaten vormen een methodologische uitdaging. De doelstelling van het onderzoek in dit proefschrift was het genereren van empirisch bewijs om de verschillende stappen binnen een systematische review van diagnostische accuratesse studies te verbeteren. We hebben specifiek gekeken naar het zoeken van studies in de literatuur, naar het verwerken van studiekwaliteit en naar de analyse van data (meta-analyse).

Hoofdstuk 1 geeft een overzicht van de uitdagingen binnen systematische reviews van diagnostische accuratesse en een overzicht van de meest recente methodologische ontwikkelingen. De methoden besproken in dit Hoofdstuk vormen een weergave van de methoden die zullen worden aanbevolen door The Cochrane Collaboration. The Cochrane Collaboration is de grootste internationale organisatie voor de ontwikkeling, het onderhoud en de promotie van systematische reviews. Vanaf oktober 2008 zal de Cochrane Database of Systematic Reviews ook reviews van diagnostische accuratesse bevatten.

Systematische reviews van diagnostische accuratesse hebben als doel alle beschikbare wetenschappelijke bewijsvoering over een specifieke test of over de vergelijking tussen meerdere tests in kaart te brengen en te evalueren. Wanneer het zoeken naar beschikbare literatuur in elektronische databases, zoals MEDLINE, teveel niet-relevante studies oplevert, kan een zoekfilter voor diagnostische studies deze overdaad inperken. Maar een mogelijk nadeel hiervan is dat studies gemist worden die wel relevant zijn. Het doel van **Hoofdstuk 2** presenteerde studie was te onderzoeken welk percentage van de relevante studies gemist wordt wanneer zoekfilters gebruikt worden. We hebben ook onderzocht in hoeverre het aantal studies dat gescreend moeten worden om één relevante studie te kunnen selecteren, vermindert door het gebruik van diagnostische zoekfilters. Het gebruik van zoekfilters voor diagnostische studies leidde tot een verlies van relevante artikelen, variërend van een gemid-

delde van 2% van het totaal aantal relevante studies dat was opgenomen in een review, tot 42%. De voornaamste redenen voor dit verlies van artikelen waren de slechte indexering van artikelen in Medline en de grote variatie in mogelijke onderzoeksopzet voor diagnostische studies. Daarnaast zorgen de zoekfilters niet gegarandeerd voor een reductie in het aantal te screenen studies, dus de impact van de filters op zoekefficiëntie zal klein zijn. Wij vinden daarom dat het gebruik zoekfilters in de ontwikkeling van een systematic review ontmoedigd moet worden.

Het doel van het onderzoek in **Hoofdstuk 3** was het vaststellen in welke mate verschillende strategieën om studiekwaliteit te definiëren en te analyseren, de resultaten van een diagnostische meta-analyse beïnvloeden. Wij heranalyseerden de data van 30 eerder gepubliceerde systematic reviews door middel van drie verschillende strategieën die zowel in de definitie van kwaliteit als in statistische benadering verschilden: (1) de analyse beperken tot de studies van hoge kwaliteit; (2) multivariabele correctie van de resultaten voor een vooraf geselecteerde set kwaliteitseisen; en (3) multivariabele correctie van de resultaten voor alleen de kwaliteitseisen die in een univariabele analyse als significant naar voren kwamen. Wij vonden geen aanwijzingen voor de hypothese dat het corrigeren voor studiekwaliteit zal leiden tot minder optimistische schattingen van diagnostische accuratesse, met minder variabiliteit in resultaten uit studies van betere kwaliteit. Het effect van iedere strategie varieerde namelijk sterk tussen de reviews, en ook waren er grote verschillen tussen de resultaten van de verschillende strategieën binnen één review.

Hoofdstuk 4 heeft betrekking op vertekening die kan ontstaan bij de evaluatie van een test die continue testresultaten geeft: het achteraf vaststellen van de optimale afkapwaarde. De optimale afkapwaarde voor de definitie van een afwijkend (positief) resultaat wordt vaak gedaan aan de hand van de in die studie verkregen resultaten. Omdat die kan leiden tot een overschatting van de accuratesse van een test (de test lijkt dus beter dan het in werkelijkheid is), hebben wij door middel van simulaties onderzocht hoe groot deze overschatting kan zijn. Verder hebben we drie alternatieve benaderingen beoordeeld op hun vermogen om deze vertekening te reduceren: (1) aannemen dat er sprake is van een bepaalde onderliggende distributie van de data; (2) leave-one-out, het telkens weglaten van een testuitslag en vervolgens bepalen of deze testuitslag op grond van de andere testuitslagen een positieve of negatieve testuitslag zou zijn geweest; en (3) de smoothed ROC methode, waarbij de geobserveerde, vaak schokkerige (zie Figuur 4.1) ROC curve op een non-parametrische manier glad getrokken wordt. De grootte van de vertekening door het achteraf vaststellen van de optimale afkapwaarde was omgekeerd evenredig aan de grootte van de studie. De onderliggende verdeling van testresultaten had weinig effect op de mate van vertekening als de grootte van de studie constant werd gehouden. Meer robuuste methoden om afkapwaarden te bepalen leidden tot minder vertekening, mits werd voldaan aan de onderliggende aannames.

Hoofdstuk 5 gaat dieper in op een mogelijke bron van heterogeniteit tussen studies: verschillen in de prevalentie van de aandoening in kwestie. Hoewel vaak beweerd

wordt dat sensitiviteit en specificiteit niet afhankelijk zijn van ziekteprevalentie, bespreken wij een aantal klinische voorbeelden waarin diagnostische accuratesse veranderde als de prevalentie ook veranderde. Gelijktijdige veranderingen in prevalentie en in sensitiviteit en specificiteit kunnen veroorzaakt worden door klinische variabiliteit of door kunstmatige variabiliteit. Klinische variabiliteit verwijst naar verschillen in de klinische situatie. Een patiëntenpopulatie met een hogere prevalentie kan bijvoorbeeld ook meer ernstig zieke patiënten omvatten, waardoor de test beter presteert. Kunstmatige variabiliteit verwijst naar effecten op prevalentie en accuratesse die geassocieerd zijn met studie opzet, bijvoorbeeld de verificatie van testresultaten door een referentiestandaard. Sensitiviteit en specificiteit zijn derhalve geen vaststaande testkarakteristieken, maar eigenschappen die de prestatie van een bepaalde test in een bepaalde situatie omschrijven. Wanneer de setting, de verwijzroute of patiëntpopulaties veranderen, kunnen ook de prevalentie en de diagnostische accuratesse van een test veranderen. Om deze reden kunnen verschillen in ziekteprevalentie en verschillen in accuratesse tussen studies fungeren als uitgangspunt voor het vinden van belangrijke verschillen in studiepopulatie of studie opzet, die de gerapporteerde accuratesse kunnen beïnvloeden.

In **Hoofdstuk 6** hebben we op systematische wijze de literatuur samengevat over de diagnostische accuratesse van fibronectine tests voor het voorspellen van zwangerschapsvergiftiging, één van de meest belangrijke oorzaken van maternale en foetale sterfte en ziekte wereldwijd. Slechts vijf studies rapporteerden voldoende data om de accuratesse te kunnen berekenen, zoals sensitiviteit en specificiteit. Voor cellulair fibronectine varieerde de specificiteit tussen de 72 en 96% als de sensitiviteit op minimaal 50% was gesteld. Voor totaal fibronectine varieerde de specificiteit onder deze omstandigheden tussen de 42 en 94%. Vanwege het geringe aantal studies en de klinische heterogeniteit tussen studies hebben we geen meta-analyse gedaan.

Hoofdstuk 7 bevat een systematische review over de diagnostische accuratesse van galactomannan detectie in serum voor de diagnose van invasieve aspergillose (IA) in immuuncompromitteerde patiënten. De meta-analyses werden uitgevoerd met 29 studies die met elkaar overeenkwamen in patiëntenpopulatie en die dezelfde criteria als referentiestandaard gebruikten. We hebben de resultaten van de meta-analyse vertaald naar een klinisch voorbeeld. In dit voorbeeld kijken we naar een afkapwaarde van 0.5 ODI (een maat voor de hoeveelheid galactomannan in serum) in een groep van 100 immuuncompromitteerde personen, met een prevalentie van IA van 8%. Dit betekent dat in deze groep 2 patiënten die IA hebben, gemist zullen worden door de test (gemiddelde sensitiviteit was 79%; dus van de 8 mensen met IA heeft 21% een fout-negatief resultaat) en dat 17 patiënten zonder IA onnodig behandeld zullen worden (gemiddelde specificiteit was 82%, dus van de 92 mensen zonder IA heeft 18% een fout-positief resultaat). Als we een afkapwaarde van 1.5 ODI hanteren in dezelfde groep patiënten dan zullen er 3 IA patiënten door de test gemist worden (sensitiviteit 62%; 38% fout-negatief) en 5 patiënten zullen onnodig behandeld worden (specificiteit 95%, 5% fout-positief). Om de toekomstige rol van

de galactomannan test voor de praktijk te bepalen, is aanvullende informatie nodig over: (1) het moment in het ziekteverloop waarop de galactomannan test positief is; (2) het moment in het ziekteverloop waarop andere tests positief zijn, zoals symptomen of CT scans; en (3) of eerder opsporen en behandelen van de ziekte ook daadwerkelijk de overleving van deze patiënten verbetert.

Discussie

Systematic reviews van de diagnostische accuratesse van een test zijn gecompliceerder dan systematic reviews van randomized controlled trials (gerandomiseerde experimenten). Dit begint al bij het formuleren van de vraagstelling, waarin de werkelijke of beoogde rol van de test in de klinische praktijk en specificaties over het patiëntenspectrum belangrijke elementen zijn. In dit proefschrift stonden de volgende drie stappen in het review proces centraal: het zoeken van literatuur, het integreren van studiekwaliteit en het analyseren van de data.

Het identificeren van studies van de diagnostische accuratesse van een test wordt bemoeilijkt door de slechte indexering van diagnostische studies in bibliografische databestanden en door de grote variatie aan mogelijke studie opzetten voor dergelijke studies. Toen The Cochrane Collaboration begon met haar Database of Systematic Reviews en met de ontwikkeling van interventie reviews, golden dezelfde problemen voor interventiestudies. Sinds die tijd is veel energie gestoken in de implementatie van een duidelijke en eenduidige indexeringsterm voor deze studies (iedere randomized controlled trial is gemerkt met het Publicatie Type “randomized controlled trial”) en de ontwikkeling van een register van randomized controlled trials en clinical trials (CENTRAL). Men kan zich afvragen of dezelfde energie en moeite gestoken dient te worden in de indexering en registratie van diagnostische studies.

Door de grote variatie in opzet en uitvoer van diagnostische accuratessestudies zal het herkennen en eenduidig labelen van deze studies in de praktijk kunnen tegenvallen. Bijvoorbeeld, data die betrekking hebben op de diagnostische accuratesse van een test kunnen verborgen zitten in studies waarvan het primaire doel niet het onderzoeken van de accuratesse was. Momenteel zijn er geen aanwijzingen dat het missen van een of twee studies in een diagnostisch review tot andere conclusies zal leiden. Zoals we eerder in Hoofdstuk 2 zagen, zijn moeilijk te achterhalen studies vaak oudere studies, waarin sowieso gedateerde testen gebruikt kunnen zijn. In veel gevallen zal het zoeken met termen voor index test(en) en aandoening volstaan.

Het is daarom mogelijk efficiënter om meer energie te stoppen in het bevorderen van informatieve rapportage in individuele studies en in een betere implementatie van het STARD statement. Complete en transparante verslaglegging zal de zichtbaarheid en dus de vindbaarheid van diagnostische accuratessestudies verhogen. Een heldere verslaglegging van studiekekenmerken is ook nodig voor een beter begrip van

de relatie tussen methodologische kwaliteit en vertekening van resultaten. Hoewel er duidelijk aanwijzingen zijn dat bepaalde studiekenmerken in individuele studies kunnen leiden tot vertekende resultaten kunnen de met elkaar verweven effecten van verschillende kwaliteitskenmerken laat zich moeilijker voorspellen. Daarbij zal het belang van de verschillende kwaliteitscriteria variëren van het ene onderzoek naar het andere. Maar ook onderzoek naar deze effecten worden gehinderd door een slechte rapportage van studiekenmerken.

Betere verslaglegging van studiekenmerken verbetert mogelijk ook het onderzoek naar andere bronnen van heterogeniteit tussen studies naast methodologische kwaliteit. Voorbeelden zijn verschillen in patiëntenspectrum, in setting of verwijspatroon van patiënten, en in de te evalueren test. In diagnostische reviews is heterogeniteit meer regel dan uitzondering, statistische methoden die hiermee rekening houden (random effect models) zijn dus de aangewezen methoden voor de analyse van deze data. Heterogeniteit kan onderzocht worden door studiekenmerken als covariabele toe te voegen in de statistische modellen, maar het trekken van conclusies uit deze modellen heeft alleen zin als er voldoende informatie wordt verschaft door de in de analyses opgenomen studies.

Het lastig kunnen vinden van studies en de slechte rapportage van methodologische kwaliteit en studiekenmerken bemoeilijken de interpretatie van de resultaten die uit systematic reviews van diagnostische accuratesse voortkomen. Gemiddelde schattingen van sensitiviteit en specificiteit alléén geven onvoldoende informatie. In de eerste plaats zijn het geen onveranderlijke eigenschappen en kunnen ze anders zijn als de omstandigheden en de patiënten verschillen. Het is echter moeilijk in te schatten in welke mate deze eigenschappen kunnen verschillen, omdat ze vertekend kunnen zijn door andere, vaak slecht gerapporteerde studiekenmerken en tekortkomingen in de methodologie. Ten tweede is een test nooit op zichzelf staand. Diagnostische testen worden ingezet om de onzekerheid over de gezondheid van een patiënt te reduceren en iedere eerder afgenomen test zal de mate waarin de volgende test deze onzekerheid vermindert, beïnvloeden. Zolang individuele studies niet naar combinaties van verschillende testen kijken, zal het lastig zijn om vergelijkende vraagstellingen (test toevoegen; test voor de voorselectie van patiënten; test vervangen) in een systematic review beperkt blijven tot indirecte vergelijkingen, waarbij ook patiëntengroepen en referentiestandaard kunnen verschillen.

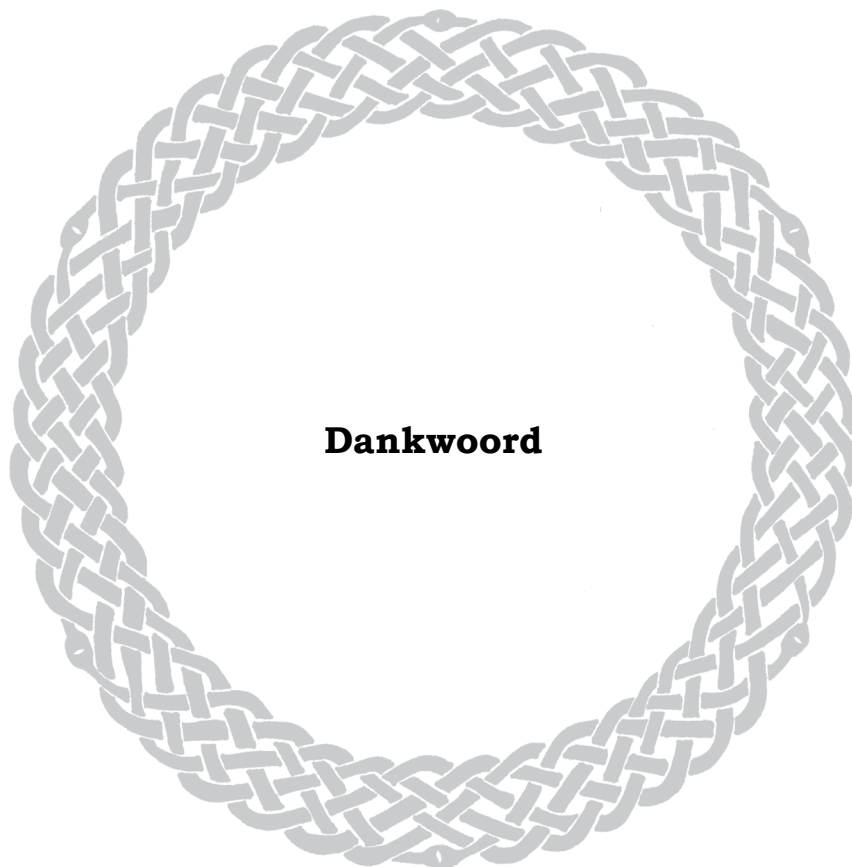
De derde reden waarom schattingen van sensitiviteit en specificiteit alleen onvoldoende informatie leveren, is omdat informatie over de gevolgen voor fout-positieve en fout-negatieve testresultaten bekend moeten zijn om het klinische nut van een test te kunnen beoordelen. In het geval van fout-positieve resultaten, zijn de volgende vragen bijvoorbeeld van belang. Worden patiënten met een fout-positief testresultaat verwezen voor verder (mogelijk invasief) onderzoek, zullen ze een redelijke goedkope en onschadelijke therapie ondergaan, of worden ze doorverwezen voor chirurgie? En om hoeveel patiënten zal het gaan op jaarbasis? In het geval van fout-negatieve resultaten willen we weten of het om een ernstige aandoening gaat.

Worden er patiënten naar huis gestuurd en niet meer gezien of zullen we deze patiënten blijven volgen, zodat we kunnen ingrijpen zodra we aanvullende informatie hebben? En wederom is het van belang te weten om hoeveel patiënten het per jaar gaat. Wanneer er onduidelijkheid is over dit soort vragen, leveren de studies die in het review opgenomen worden mogelijk informatie over wat gangbaar is in de klinische praktijk en wat de beoogde rol van de geëvalueerde test in deze situaties kan zijn.

We kunnen concluderen dat de ontwikkeling van de methoden voor het vinden van literatuur, voor het vaststellen van de methodologische kwaliteit, voor de analyse van de data, en voor het onderzoeken van statistische en klinische heterogeniteit, baat zullen hebben bij een beter verslaglegging van de opzet en kenmerken van diagnostische studies. Hoewel kwaliteit van rapportage niet hetzelfde is als methodologische kwaliteit van een studie, zal een betere rapportage van studiekenmerken wel de interpretatie van studieresultaten en daarmee de algehele kwaliteit van een studie bevorderen.

De ontwikkeling en uitvoering van systematische reviews van diagnostische accuratesse is lastig en de methodologie ervoor is nog steeds in ontwikkeling. Het vertalen van de resultaten naar beleid en praktijk is een zo mogelijk nog grotere uitdaging. Het vraagt kennis van zowel de methodologie van diagnostisch onderzoek als kennis van de klinische context waarin de test of testen gebruikt zullen worden. Als de rapportage van diagnostisch onderzoek verbetert, zouden we auteurs van systematische reviews van diagnostische accuratesse willen vragen hun lezers aan de hand te nemen en hen te begeleiden in het begrijpen en vertalen van de gevolgen van hun resultaten. Alleen dan kunnen zij gerust zijn dat deze resultaten hun weg zullen vinden naar de klinische praktijk en dat zij de zorg voor patiënten zullen verbeteren.





Dankwoord

De meest gestelde vraag in de afgelopen vier jaar was de vraag wat een dierenarts in het AMC te zoeken heeft, in een project dat niets met dieren van doen heeft. Het antwoord op die vraag voert mij steevast naar het eerste jaar diergeneeskunde, waar een bevlogen hoogleraar een vlammend betoog hield over het belang van dierziekten voor de gezondheid van mensen. Toen ik een aantal jaren later onderzoeksstage ging doen bij diezelfde hoogleraar, werd ik gegrepen door het fenomeen 'onderzoek'. Het iedere keer weer opnieuw geconfronteerd worden met nieuwe vragen die de nieuwsgierigheid prikkelen en dan ook nog de mogelijkheid hebben om op zoek te gaan naar de antwoorden. Ik begon mij dan ook af te vragen of ik niet meer in de wieg was gelegd voor onderzoek dan voor praktijk. Frans van Knapen en Ad Koets, zonder jullie goede raad aan het eind van die vijftien maanden was ik waarschijnlijk niet zo snel in het onderzoek terecht gekomen. En Frans, zonder jouw soms overdonderende enthousiasme en je rotsvaste geloof in het feit dat een dierenarts er niet alleen is om zieke dieren beter te maken, was het AMC als mogelijke werkgever nooit in mijn hoofd opgekomen.

Deze dierenarts ging dus solliciteren bij de afdeling Klinische Epidemiologie, Biostatistiek en Bioinformatica (KEBB). Toen ik erachter kwam dat ik vakken als graslandbeheer en een cursus 'verdoven op afstand' nog op mijn CV had staan, wist ik eigenlijk al zeker dat ik nooit de meest geschikte kandidaat zou kunnen zijn. Maar dat pakte gelukkig anders uit. Patrick Bossuyt, Hans Reitsma en Rob Scholten, dank dat jullie vertrouwen in mij stelden en je niet lieten afschrikken door allerlei rare keuzevakken. Pas nu, tijdens het schrijven van het dankwoord en het terugblikken op de afgelopen vier jaar besef ik wat een geweldige tijd ik heb gehad.

Patrick, je scherpzinnigheid en je brede kennis zijn al in ontelbare proefschriften geroemd. Maar wat ik vooral waardeer is dat je me enorm betrokken bij alles wat er op het gebied van diagnostische reviews en diagnostiek van infectieziekten gebeurde. Op die manier heb je me de mogelijkheid gegeven veel van de wereld te zien en was ik steeds op de hoogte van de meest recente ontwikkelingen. Hans, ondanks dat je de enige echte GVR (Grote Vriendelijke Reus) bent, ben jij ook in staat om tot wanhoop drijvende vragen te stellen. Jij wist mij met je vele vragen en ideeën uit te dagen en mijn nieuwsgierigheid te prikkelen. Rob, over jou kan ik kort zijn. Als er een verkiezing van beste dagelijks begeleider van het jaar is, dan ga ik je daar zeker voor nomineren.

Zelfs de leukste baan wordt een saaie bedoening zonder gezellige collega's. Ook daar ontbrak het me de afgelopen jaren niet aan, de KEBB omvat immers een breed scala aan mensen met verschillende achtergronden. De lunches en borrels waren altijd ontzettend gezellig met (ex)promovendi Anouk, Barbara v M, Bart, Hans W, Helene, Iris, Jeroen, Joost, Kim, Kimberley, Marije, Marjolein, Nadine, Olga, Rebecca, Sandra, Susanne, en Willem.

Mijn speciale dank gaat uit naar mijn (ex)kamerogenoten, Nynke, Marlies, Fleur en Teodora. Teodora, heel veel succes met je promotie, ik hoop dat het alleen maar beter wordt. Nynke, eigenlijk zit je nu half op een andere afdeling, maar gelukkig hebben we daar nog niet veel van gemerkt. Veel geluk met je verdere loopbaan. Met jouw nuchtere ‘down to earth’ instelling zit dat wel goed. Fleur, met jouw komst naar onze kamer bracht je een hoop drukte en gezelligheid met je mee. Ik ben je vrolijkheid en je sociale instelling enorm gaan waarderen en blijf graag nog een tijdje bij jou op de kamer zitten. Marlies, we zijn op dezelfde dag begonnen en gaan straks met een week tussentijd ons proefschrift verdedigen. Ik vond (en vind) je een fijne collega, waardoor het helemaal niet erg was als onze namen weer eens werden verwisseld. Succes straks!

Om te kunnen promoveren, moet er wel wat geschreven worden. Dat gaat niet in je eentje. Anne Rutjes en Marcello Di Nisio, ik ben jullie veel dank verschuldigd voor de dataset waar ik wel de lusten, maar niet de lasten van mocht hebben. Carl Moons, jullie hebben in Utrecht toch wat andere ideeën over diagnostiek dan wij hier in Amsterdam. Maar zonder discussie geen vooruitgang. Ik waardeer je inbreng in ons cut-off stuk. Koos Zwinderman, vooral je vermogen om de meest ingewikkelde formules zodanig uit te leggen dat de toehoorders zelf gaan geloven dat het echt niet ingewikkelder is dan eenvoudig optellen en aftrekken, doet me iedere keer weer versteld staan.

Jeltsje, Joris, Ben Willem en Gerben, ik vond het geweldig om deel uit te mogen maken van jullie team. Jeltsje, ik hoop dat ik je toch een beetje heb kunnen helpen, ondanks het feit dat jullie vaak tegen problemen aan liepen op het moment dat er nog geen oplossingen voor waren. Ben Willem, dank ook dat je mijn manuscript hebt willen beoordelen. Khalid, I enjoyed the dinner and the discussions at the (almost) end of the pre-eclampsia project.

Yvette, Caroline, Henk en Christina, het team van microbiologen die mij hebben geholpen bij het schrijven van de pilot review. Het was veel werk, maar ik heb ook veel van jullie kunnen leren. Christina, bedankt ook voor je bereidheid om voor ons op zoek te gaan naar een geschikt onderwerp. Nynke Smidt, onder andere dank voor de tijd die je gestopt hebt in het managen van de pilot reviews (waaronder deze).

Chapter 1 of this thesis would not have existed without Jon Deeks and Constantine Gatsonis. Moreover, without the members of the Screening and Diagnostic Test Methods Group, there would perhaps not have been a thesis like this at all. Jon, I was honoured to be your ‘paranimf’, thank you for that, for being such a kind person and good teacher (I love the teddy bear slides), and for the interesting discussions about diagnostics. Les Irwig, I enjoyed our teleconferences. Those discussions taught me to get more grip on something as abstract as diagnostic test accuracy.

Leden van de promotiecommissie, prof.dr. Knottnerus, prof.dr. Offringa, prof.dr. Speelman, prof.dr. T. Stijnen en prof. dr. De Vet, hartelijk dank voor het beoordelen van de inhoud van dit proefschrift. Ik zie echt uit naar de verdediging.

Als het dan bijna zover is en de promotie is in zicht, zijn er nog een paar mensen die van onschatbare waarde zijn voor het geluk van de promovenda. Allereerst zijn daar Gertie en Ferdinand, die de ruwe versie op tikfouten, rare verwijzingen en kromme zinsconstructies hebben gecontroleerd. En Ferdinand, het is daarbij niet de bedoeling dat je dan nog hele kritische vragen gaat stellen en vraagtekens zet bij de inhoud. Maar gelukkig weet je dat ik dat alleen maar heel erg kan waarderen.

Zonder 'echt' boekje geen verdediging. Koen, ik waardeer je principes, je duidelijkheid en vooral je vriendschap. Ik vind het geweldig dat je voor mij de opmaak wilde doen. Maar voorlopig kan ik even geen pizza's en koffie meer zien.

Dan zijn er nog Gré en Petra, van het KEBB-secretariaat. Altijd bereid om je te helpen als je er zelf even niet uitkomt. In het heetst van de strijd staan zij voor je klaar met goede raad en daad. Bang dat je toch een pagina verkeerd gekopieerd hebt? Geen nood meid, dan halen we alle enveloppen toch weer open en beginnen we gewoon opnieuw. En nu we het toch over het secretariaat hebben, mag ik Hanni natuurlijk niet vergeten. Hanni, als ik jou niet had, dan zouden er heel wat SR-cursusdagen door mijn toedoen in de soep zijn gelopen. Ik ben blij dat jij zo georganiseerd bent.

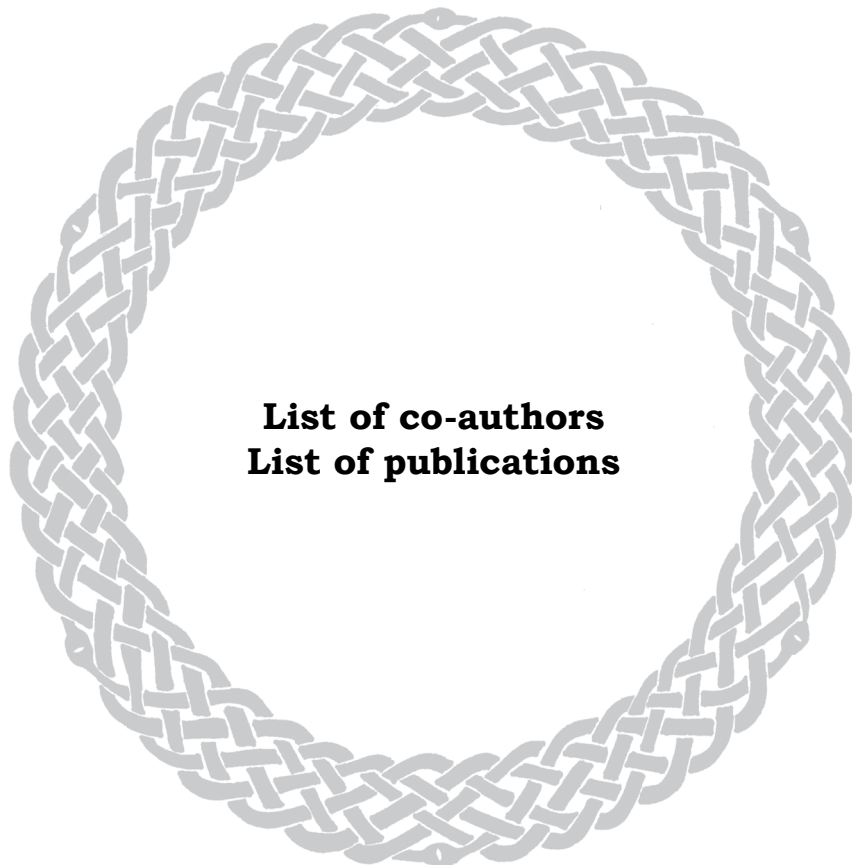
Lotty en Janneke, mijn paranimfen, jullie zullen straks mijn steun en toeverlaat zijn. Maar eigenlijk zijn jullie dat nu al. Lotty, ik heb jou leren kennen als een heel attent persoon. Jij biedt iedereen een luisterend oor en zal nooit een verjaardag of andere belangrijke gebeurtenis in iemands leven vergeten. Hoe doe je dat toch? Janne, met jou kan ik lekker ouwehoeren over DIO, problemen met diagnostiek in de praktijk en over helemaal niks. Jij bent iemand die een ander nooit zal beoordelen op uiterlijk of rare gewoonten. Weet dat onze deur altijd voor je open staat.

Ralph, jij bent de vrolijke noot in mijn leven. Maar daarnaast ben jij ook degene die beter voor me zorgt dan wie dan ook. Jij ving me bijna letterlijk op toen ik viel. Zonder jou was ik allang in zeven sloten tegelijk gelopen zonder er ooit weer uit te komen.

Tot slot een woord van dank voor mijn vader en moeder, Klaas en Gré Leeftang. Alles wat ik ben, ben ik dankzij jullie. Pa, van jou heb ik duidelijk je bescheidenheid, je onzekerheid en je absolute bereidheid iets goeds te doen voor deze wereld. Jij hebt het vrijwilligerswerk en de liefde voor de natuur er met de paplepel bij ons ingegoten. Ma, van jou komen de nuchterheid, de oprechtheid en de vrolijkheid die alles weer in balans brengen. Jullie hebben Arjan en mij altijd vrij gelaten te doen wat we wilden en dat is van onschatbare waarde.

Dankwoord





List of co-authors
List of publications

List of co-authors

Henk A. Bijlmer. Department of Clinical Microbiology and Infection Control, Bronovo hospital, The Hague, The Netherlands.

Patrick M.M. Bossuyt. Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands.

Jeltsje S. Cnossen. Department of General Practice, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands.

Yvette J. Debets-Ossenkopp. Department of Clinical Microbiology and Infection Control, VU Medical Centre, Free University, Amsterdam, The Netherlands.

Jon J. Deeks. Department of Public Health and Epidemiology, University of Birmingham, United Kingdom.

Marcello Di Nisio. Department of Medicine and Aging, School of Medicine and Aging Research Center, Ce.S.I., "Gabriele D'Annunzio" University Foundation, Chieti-Pescara, Italy.

Constantine Gatsonis. Center for Statistical Sciences, Brown University, Providence, USA.

Lotty Hooft. Dutch Cochrane Centre, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands.

Les Irwig. Screening and Test Evaluation Program, School of Public Health, University of Sydney, Sydney, Australia

Khalid S. Khan. Department of Obstetrics and Gynaecology, Birmingham Women's Hospital, Birmingham, UK.

Ben Willem Mol. Department of Obstetrics and Gynecology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Karel G.M. Moons. Julius Center for Health Sciences and General Practice, University Medical Center, Utrecht, The Netherlands.

Joris A. van der Post. Department of Obstetrics and Gynecology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Johannes B. Reitsma. Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands.

Gerben ter Riet. Horten Center, University of Zurich, Zurich, Switzerland.

Anne W.S. Rutjes. Department of Clinical Pharmacology and Epidemiology, Consorzio Mario Negri Sud, Chieti, Italy.

Rob J.P.M. Scholten. The Dutch Cochrane Centre, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Christina M.J.E. Vandenbroucke-Grauls. Department of Clinical Microbiology and Infection Control, VU Medical Centre, Free University, Amsterdam, The Netherlands.

Caroline E. Visser. Department of Medical Microbiology, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands.

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

List of publications

Meads CA, Cnossen JS, Meher S, Juarez-Garcia A, Ter Riet G, Duley L, Roberts TE, Mol BW, van der Post JA, **Leeflang MM**, Barton PM, Hyde CJ, Gupta JK, Khan KS. Methods of prediction and prevention of pre-eclampsia: systematic reviews of accuracy and effectiveness literature with economic modelling. *Health Technol Assess*. 2008;12(6):1-270.

Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clin Chem*. 2008;54(4):729-37.

Brenninkmeijer EE, Schram ME, **Leeflang MM**, Bos JD, Spuls PI. Diagnostic criteria for atopic dermatitis: a systematic review. *Br J Dermatol*. 2008;158(4):754-65.

Cnossen JS, **Leeflang MM**, de Haan EE, Mol BW, van der Post JA, Khan KS, ter Riet G. Accuracy of body mass index in predicting pre-eclampsia: bivariate meta-analysis. *BJOG*. 2007;114(12):1477-85.

Leeflang MM, Cnossen JS, van der Post JA, Mol BW, Khan KS, ter Riet G. Accuracy of fibronectin tests for the prediction of pre-eclampsia: a systematic review. *Eur J Obstet Gynecol Reprod Biol*. 2007;133(1):12-9.

Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deeks J, Bossuyt P. Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. *Clin Chem*. 2007;53(2):164-72.

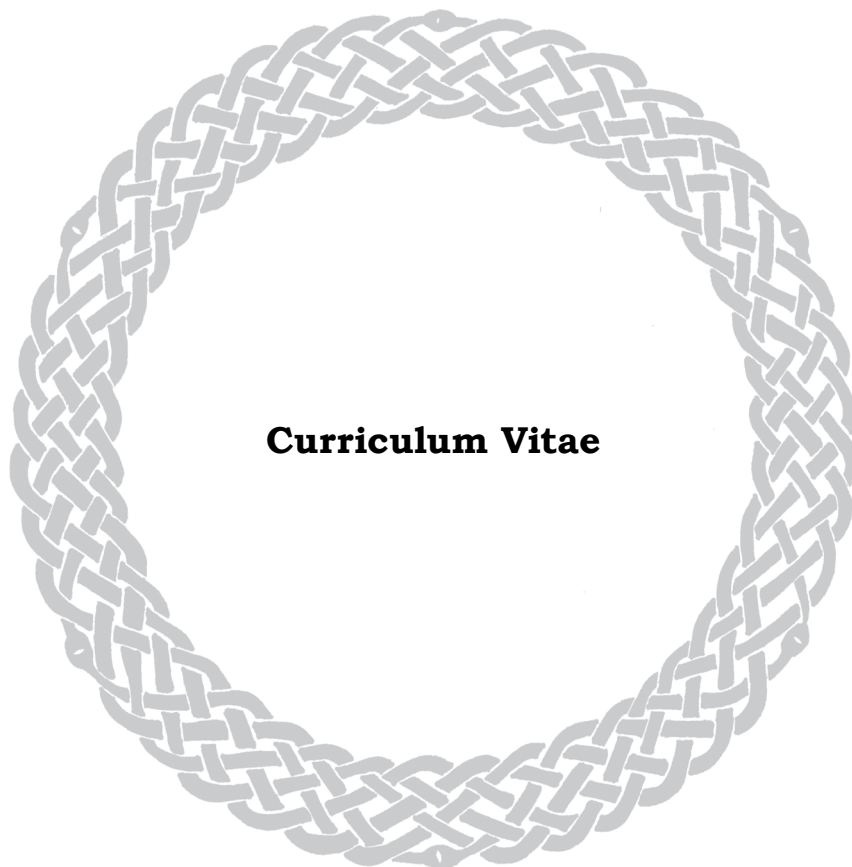
Leeflang MM, Debets-Ossenkopp YJ, Visser CE, Bossuyt PM. Meta-analysis of diagnostic test accuracy. *Clin Infect Dis*. 2006;43(9):1220.

Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *J Clin Epidemiol*. 2006;59(3):234-40.

Leeflang MM, Bossuyt PM. Test accuracy is likely to vary depending on the population it is used in. *Vet Parasitol*. 2005;134(1-2):189.

Leeflang M. Evidence-based medicine in de diergeneeskundige praktijk. *Tijdschr Diergeneeskd*. 2005;130(2):48-9.





Curriculum Vitae

Curriculum Vitae

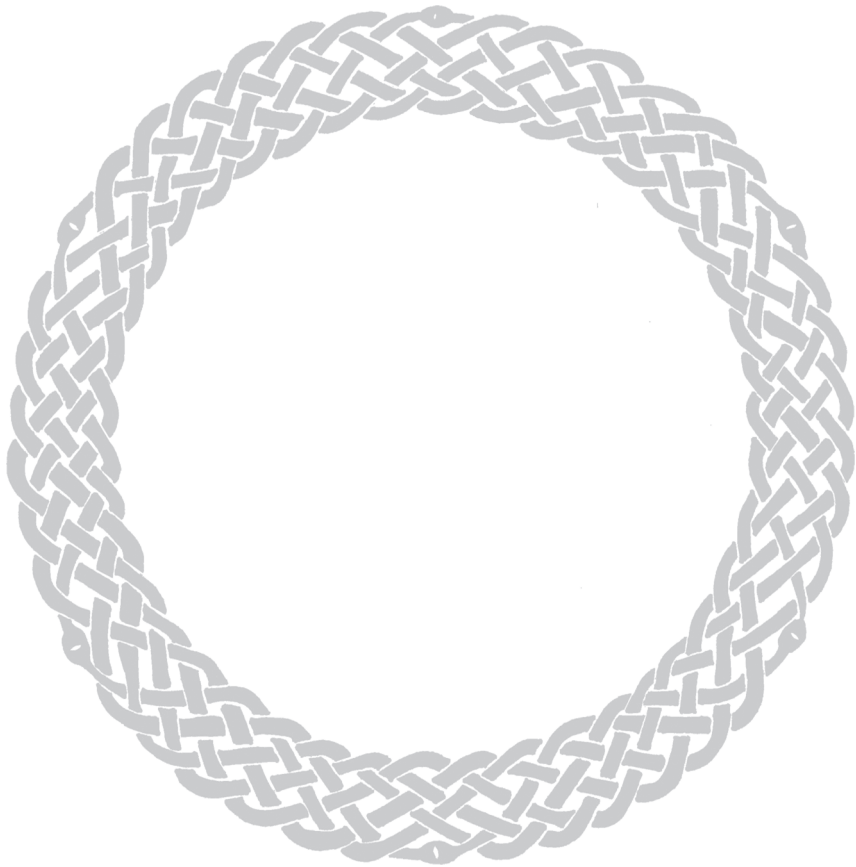
De oouevaar bracht Maria Mariska Geertruida Leeftang op 20 januari 1976 naar de Pieter Keukenstraat in Volendam. De altijd lachende peuter werd al snel een betrokken tiener, die samen met haar vriendinnetje MenS oprichtte, een tweemans actiegroep die sloten schoonmaakte, actie voerde tegen openbare verbranding van kerstbomen en geld inzamelde voor het redden van olifanten en tijgers.

Mariska doorliep het VWO op het Don Bosco College te Volendam en wilde daarna diergeneeskunde studeren in Utrecht. Na de eerste keer te zijn uitgeloot voor deze studie, was het de tweede keer dan toch raak. Er werd nog even kort getwijfeld, omdat de studie biologie toch interessanter bleek te zijn dan verwacht, maar de droom om dierenarts te worden was op dat moment toch te sterk. Tijdens het eerste jaar in Utrecht ging Mariska samenwonen met Ralph en kochten zij samen een huisje in het centrum van Volendam. De studie verliep ondanks het heen en weer gereis en allerlei bijbaantjes voorspoedig en Mariska kreeg de kans om een Excellent Tracé onderzoeksstage te doen bij de (toenmalige) afdeling Voedingsmiddelen Van Dierlijke Oorsprong en de afdeling Immunologie. Deze onderzoeksstage richtte zich op het ontwikkelen van een nieuwe detectiemethode voor *Mycobacterium avium* ssp. *paratuberculosis*. Behalve voor studeren, was er in Utrecht ook nog ruimte voor nevenactiviteiten, vooral toen de verhuizing van Volendam naar Utrecht een feit was. In 2000 begon Mariska als vrijwilliger bij Stichting Diergeneeskunde In Ontwikkelingssamenwerking (DIO). Van deze organisatie was zij van november 2002 tot november 2006 voorzitter.

Na haar afstuderen eind 2003 vond Mariska een promotieplaats aan de afdeling Klinische Epidemiologie, Biostatistiek en Bioinformatica van het Academisch Medisch Centrum in Amsterdam. Naast het onderzoek zoals dat gepresenteerd wordt in dit proefschrift, heeft zij ook verschillende onderwijstaken vervuld voor het Dutch Cochrane Centre. Vanaf februari 2007 maakt Mariska deel uit van de Continental Europe Support Unit (CESU), een eenheid ter ondersteuning van de implementatie van systematic reviews van diagnostische accuratesse binnen The Cochrane Collaboration. En vanaf april 2008 is zij twee dagen per week werkzaam voor het Koninklijk Instituut voor de Tropen bij de afdeling biomedical research.









Stellingen

Behorende bij het proefschrift

Systematic Reviews of Diagnostic Test Accuracy

1. Methodologische zoekfilters maken het zoekproces van een systematic review niet efficiënter. (*dit proefschrift, hoofdstuk 2*)
2. Correctie voor kwaliteit in meta-analyses geeft onvoorspelbare resultaten. (*dit proefschrift, hoofdstuk 3*)
3. Als de afkapwaarde voor een continue test op basis van in het onderzoek gevonden waarden wordt bepaald, leidt dit doorgaans tot een overschatting van sensitiviteit en specificiteit. (*dit proefschrift, hoofdstuk 4*)
4. Verdere studie van het verband tussen de kwaliteit van een onderzoek en mogelijke vertekening in de resultaten is zinloos als de rapportage van onderzoek niet verbetert. (*dit proefschrift, hoofdstuk 3*)
5. Zolang diagnostisch onderzoek gebrekkig wordt gerapporteerd, blijft de prevalentie van een aandoening een belangrijke indicator voor de representativiteit van het onderzoek. (*dit proefschrift, hoofdstuk 5*)
6. De regel van Bayes dient met voorzichtigheid geïnterpreteerd te worden. (*dit proefschrift, hoofdstuk 5*)
7. De controverse over wat klinici beter begrijpen, likelihood ratio's dan wel sensitiviteit en specificiteit, is een non-discussie.
8. Het negeren van wetenschappelijke onderbouwing ('evidence') in de klinische praktijk is een verspilling van tijd, geld en moeite, maar is vooral ook onethisch.
9. Between animal and human medicine there is no dividing line, nor should there be. The object is different but the experience obtained constitutes the basis of all medicine. (*Rudolf Virchow, arts, antropoloog en politicus, 1821-1902*)
10. Een ziekenhuis laat zich goed vergelijken met een varkensstal.

Mariska Leeflang, Amsterdam, 1 juli 2008