Beyond diagnostic accuracy. Applying and extending methods for diagnostic test research
Glas, A.S.

Citation for published version (APA):

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: http://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.
Beyond diagnostic accuracy
Applying and Extending Methods for
Diagnostic Test Research

Glas, Afina Siberta
All published papers and illustrations were reprinted by permission of the copyright owner.
© 2003 Afina S. Glas, Amsterdam, The Netherlands

ISBN 90-9016966-0

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic, or mechanical, including photocopy, recording or any information storage and retrieval system, without written permission of the copyright owner.

Cover: Roel Stavorinus
Layout: Roel Stavorinus
Printed by: Thela Thesis, Rozengracht 176A,
          1016 NK Amsterdam.
          info@thelathesis.nl

The financial support by the following institutes and companies for the publication of this thesis is gratefully acknowledged:

The University of Amsterdam
Zambon Nederland BV
Yamanouchi Pharma BV
Sanofi-Synthelabo BV
Hitachi Medical Systems BV
Beyond Diagnostic Accuracy

Applying and Extending Methods for Diagnostic Test Research

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. mr. P.F. van der Heijden
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 27 juni 2003, te 12:00 uur.

door Afina Siberta Glas

geboren te Dokkum
Promotor: Prof. Dr. P.M.M. Bossuyt
Co-promotor: Prof. Dr. A.H. Zwinderman

Faculteit der Geneeskunde, Universiteit van Amsterdam

Thesis Committee

G.J. Bonsel, M.D., PhD, professor in Social Medicine, Amsterdam Medical Center
H.R. Büller, M.D., PhD, professor of Vascular Medicine, Amsterdam Medical Center
Y. van der Graaf, M.D., PhD, professor of Clinical Epidemiology, University Medical Center, Utrecht
L.A.L.M. Kiemeney, PhD, associate professor in Epidemiology, University Medical Center St. Radboud, Nijmegen
J.A. Knottnerus, M.D., PhD, professor in General Practice, Maastricht University
J.S. Laméris, professor of Radiology, Amsterdam Medical Center
Contents

General Introduction

Part One : Methodology

1 The Diagnostic Odds Ratio. A single indicator of test performance 13

2 The sensitivity and specificity of tests that locate multiple lesions; evaluating colonography 29

3 Direct pooling of sensitivity and specificity using bivariate models in meta-analysis of studies of diagnostic accuracy 41

Part Two : Methodology in Practice

4 CT colonography at ultra-low radiation dose 61

5 Tumor markers in the diagnosis of primary bladder cancer. A systematic review. 67

6 The added value of pre-discharge Dobutamine Stress Echocardiography in the long-term risk stratification of chest pain patients 85

7 Comparison of diagnostic decision rules and structured data collection in assessment of acute ankle injury. 97

8 Physicians’ interpretation of ventilation-perfusion lung scans. 111

9 Computed Tomography versus Magnetic Resonance Imaging in the reduction of diagnostic uncertainty in patients with lumboradicular syndrome. 127

General discussion 137
Dutch summary (Samenvatting) 145
List of Publications 149
List of Co-authors 151
Words of Gratitude (Dankwoord) 155
About the Author 159
The current demand for evidence based practice generates the need for proper evaluations of the effectiveness of diagnostic and therapeutic interventions. Diagnostic effectiveness reflects the performance of diagnostic tests under ordinarily, rather than ideal conditions. Within diagnostic effectiveness studies different levels of test evaluations can be distinguished, ranging from the technical development up to cost-effectiveness.

The most frequently performed type of study is the evaluation of a test’s diagnostic accuracy. In such evaluations the results of a new test are compared with the results of a reference standard - formerly called the gold standard - which is the best available method to reveal the true disease state of the patients. When both the true disease state and the outcome of the test are dichotomous, to be classified as positive or negative, the results can be structured in a 2x2 contingency table.

Several measures of diagnostic accuracy can be calculated from such a table. The best known are the sensitivity and specificity, which stand for the proportion of positive test results within patients with the target disease and the proportion of negative test results within patients without the target disease, respectively. Other measures of diagnostic accuracy are the positive and negative predictive values, the likelihood ratios of a positive and negative test result, the overall accuracy and the Youden index. When the test outcome is continuous, while the true disease state is dichotomous, the likelihood ratios of the respective test results, the receiver operating characteristic curve and the area under this curve can be used. All these measures express the correspondence between the results of the test and the outcome of the reference standard, and can be seen as expressions of how well the test is able to distinguish between patients with the target disease from those without it.

This thesis explores the characteristics of a measure beyond the above mentioned ones: the diagnostic odds ratio (chapter one). It also provides more detailed information on the definition, interpretation and use of the outcome measures relative to the odds ratio. Yet, the reduction of the full test information into a 2x2 table, and the subsequent calculation of these measures may not be straightforward. Chapter two illustrates a number of pitfalls in studies that evaluate tests that can detect and localise specific lesions and try to force the resulting data in a 2x2 table. We present an suitable method to arrive at valid sensitivity and specificity. This method is illustrated in chapter four. In the study reported there, sensitivity and specificity were estimated for CT colonography in the detection of polyps in the colon at different levels of radiation dose.

When one wants to arrive at valid and precise estimates of a test’s worth, one
can review the literature for available evidence on the accuracy of a diagnostic test. If one uses an explicit and complete search and includes criteria for evaluating the potential for bias in the studies found, the exercise is called a systematic review. A researcher who wants to do a systematic review aims at identifying all studies that have compared the test of interest with the reference standard. If enough comparable studies are available, a meta-analysis can be performed to arrive at summary measures of test performance.

A well known difficulty in meta-analysis of diagnostic tests is the existence of considerable variation between the results from different studies. A part of that variability can be attributed to differences in the threshold used to classify a test result as positive or negative. Simple averaging the results, with only adjusting for the sample size, may then yield invalid results. The most widely used class of methods for meta-analysis of diagnostic accuracy studies is based on the concept of the summary receiver operating characteristic (SROC) curve, which can take this threshold effect into account. This curve is a global measure of the accuracy of the test, which however has an ambiguous interpretation, and which does not allow for straightforward calculation of standard errors. Chapter three introduces a bivariate approach as an alternative to estimate summary values of the well established measures sensitivity and specificity. This method can be extended to a meta-regression technique to explore sources of between-study variability separately on sensitivity and specificity. Chapter five contains an application of this bivariate approach to the meta-analysis of diagnostic tests in a systematic review of urine based tumor markers in the primary diagnosis of bladder cancer.

All measures of diagnostic accuracy share a disadvantage, they do not indicate what the test's added value is relative to what is already known about the patient. The added value of a diagnostic test depends by definition on the information available before testing, in particular on the likelihood of (non)disease. Chapter six of this thesis reports on a study of the added value of the dobutamine stress echocardiography in the diagnosis of cardiac disease in patients at known low risk. This test has been shown to be associated with the cardiac prognosis of patients discharged from an emergency department. At discharge, physicians will have a prognostic idea about these patients. We invited these physicians to express their prognostic guess as (subjective) probability estimates without knowing the test result and evaluated the contribution of dobutamine stress echocardiography. The test results can be linked to the pre-test probability estimates using likelihood ratios and Bayes theorem.

It is unclear as to what extent subjective probability expressions assigned by physicians can be relied on. Chapter seven presents a comparison of subjective
probabilities and two formal diagnostic decision rules in the diagnosis of ankle fracture in patients with a fresh ankle sprain. In chapter eight, an evaluation is described of physicians’ subjective judgment in interpreting the ventilation/perfusion lung scan in the diagnosis of pulmonary embolism. We used the subjective pre-test and post-test probability estimates of the lung scan to derive subjective likelihood ratios for the respective test results. These subjective likelihood ratios were then compared to the objective likelihood ratios of the lung scan, as obtained from the comparison of the test results with the outcome of the reference standard.

In general the difference between pre-test and post-test probabilities can be used to characterize the information content of a diagnostic test. Chapter nine presents an example of how the information content can be used to compare computer tomography with magnetic resonance imaging in the diagnosis of patients with lumboradicular syndrome.

Tests have to be properly evaluated to see whether or not they are able to improve or maintain the health status of patients. This thesis is based on both existing methods and an extension of the methodology used to evaluate diagnostic tests. In the final chapter we summarize the state of the art of diagnostic methodology and provide indications for future research.

References
1 The Diagnostic Odds Ratio: A single indicator of test performance

Abstract
Diagnostic testing can be used to discriminate subjects with a target disorder from subjects without it. Several indicators of diagnostic performance have been proposed, such as sensitivity and specificity. Using paired indicators can be a disadvantage in comparing the performance of competing tests, especially if one test does not outperform the other on both indicators. Here we propose the use of the odds ratio as a single indicator of diagnostic performance. The diagnostic odds ratio is closely linked to existing indicators, it facilitates formal meta-analysis of studies on diagnostic test performance, and it is derived from logistic models, which allow for the inclusion of additional variables to correct for heterogeneity. A disadvantage is the impossibility of weighing the true positive-and false positive rate separately. In this paper the application of the diagnostic odds ratio in test evaluation is illustrated.
1 The Diagnostic Odds Ratio: A single indicator of test performance

Introduction
In an era of evidence-based medicine, decision makers need high-quality data to support decisions about whether or not to use a diagnostic test in a specific clinical situation and, if so, which test. Many quantitative indicators of test performance have been introduced, comprising sensitivity and specificity, predictive values, chance-corrected measures of agreement, likelihood ratios, area under the receiver operating characteristic curve, and many more. All are quantitative indicators of the test’s ability to discriminate patients with the target condition (usually the disease of interest) from those without it, resulting from a comparison of the test’s results with those from the reference standard in a series of representative patients. In most applications, the reference standard is the best available method to decide on the presence or absence of the target condition. Less well known is the odds ratio as a single indicator of test performance. The odds ratio is a familiar statistic in epidemiology expressing the strength of association between exposure and disease. As such it can also be applied to express the strength of the association between test result and disease.

This paper offers an introduction to the understanding and use of the odds ratio in diagnostic applications. In brief, we will refer to it as the diagnostic odds ratio (DOR). First we will point out the usefulness of the odds ratio in dichotomous and polychotomous tests. We will then discuss the use of the DOR in meta-analysis and the application of conditional logistic regression techniques to enhance the information resulting from such analysis.

Dichotomous test outcomes
Although most diagnostic tests have multiple or continuous outcomes, either grouping of categories or application of a cut-off value is frequently applied to classify results into positive or negative. Such a dichotomization enables to represent the comparison between a diagnostic test and its reference standard in one 2x2 contingency table, as depicted in table 1.

Table 1 2x2 contingency table.

<table>
<thead>
<tr>
<th>Reference Test</th>
<th>Target disorder</th>
<th>No target disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>test</td>
<td>positive TP</td>
<td>FP</td>
</tr>
<tr>
<td>negative</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

The abbreviations TP, FP, FN and TN denote the number of respectively true positives, false positives, false negatives and true negatives. Same definitions are used throughout the text and table 2.
Common indicators of test performance derived from such a 2x2 table are the sensitivity of the test, its specificity, the positive and negative predictive values and the positive and negative likelihood ratios.\(^1\) See table 2 for a definition of these indicators.

Unfortunately, none of these indicators in itself validly represent the test’s discriminatory performance. Sensitivity is only part of the discriminatory evidence as high sensitivity may be accompanied by low specificity. Additionally, no simple aggregation rule exists to combine sensitivity and specificity into one measure of performance.

Table 3 shows the performance of three different radiological diagnostic tests to stage ovarian cancer as an illustration of the need for combined judgment. All three tests were performed in a group of 280 patients suspected of ovarian cancer.\(^2\) Surgical and histo-pathological findings were used as the reference standard. The sensitivity of the ultrasound was worse than that of the computer tomography (CT) scan in detecting peritoneal metastases but for the specificity the reverse held. Likelihood ratios and the predictive values are also not decisive. Also the combined evidence of the pairs of indicators cannot simply be ranked.

For this, a single indicator of test performance like the test’s accuracy is required. In addition to its global meaning of agreement between test and reference standard, accuracy in its specific sense refers to the percentage of patients correctly classified by the test under evaluation. This percentage depends on the prevalence of the target disorder in the study group whenever sensitivity and specificity are not equal and it weights false positive and false negative findings equally. Another single indicator is Youden’s index.\(^3\)\(^4\) It can be derived from sensitivity and specificity and as such it is independent of prevalence, but since it is a linear transformation of the mean sensitivity and specificity its values are difficult to interpret.\(^5\)
The Diagnostic Odds Ratio: A single indicator of test performance

<table>
<thead>
<tr>
<th>Test indicator</th>
<th>formula</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (true positive rate, TPR)</td>
<td>TP/(TP+FN)</td>
<td>proportion positive test results among diseased</td>
</tr>
<tr>
<td>Specificity (true negative rate, TNR)</td>
<td>TN/(TN+FP)</td>
<td>proportion negative test results among the ‘healthy’</td>
</tr>
<tr>
<td>Positive predictive value (PPV)</td>
<td>TP/(TP+FP)</td>
<td>proportion diseased among subjects with a positive test result</td>
</tr>
<tr>
<td>Negative predictive value (NPV)</td>
<td>TN/(TN+FN)</td>
<td>proportion non-diseased among subjects with a negative test result</td>
</tr>
<tr>
<td>Likelihood ratio of a positive test result (LR+)</td>
<td>Sensitivity / (1-specificity)</td>
<td>ratio of a positive test result among diseased to the same result in the ‘healthy’</td>
</tr>
<tr>
<td>Likelihood ratio of a negative test result (LR-)</td>
<td>(1-sensitivity) / specificity</td>
<td>ratio of a negative test result among diseased to the same result in the ‘healthy’</td>
</tr>
<tr>
<td>Accuracy</td>
<td>(TP+TN)/(TP+TN+FP+FN)</td>
<td>proportion correctly identified subjects</td>
</tr>
<tr>
<td>Youden's index</td>
<td>sensitivity + specificity -1</td>
<td></td>
</tr>
</tbody>
</table>
The odds ratio used as single indicator of test performance is a third option. It is not prevalence dependent and may be easier to understand, as it is a familiar epidemiological measure. The diagnostic odds ratio of a test is the ratio of the odds of positivity in disease relative to the odds of positivity in the non-diseased.\textsuperscript{6,7} This means that the following relations hold:

\begin{equation}
DOR = \frac{TP}{FP} = \frac{sens}{(1 - sens)} \cdot \frac{(1 - spec)}{spec}
\end{equation}

Alternatively, the DOR can be read as the ratio of the odds of disease in test positives relative to the odds of disease in test negatives.

\begin{equation}
DOR = \frac{TP}{FP} = \frac{PPV}{(1 - PPV)} = \frac{LR(+)}{NPV}
\end{equation}

There is also a close relation between the DOR and likelihood ratios:

\begin{equation}
DOR = \frac{TP}{FP} = \frac{LR(\text{spec})}{LR(\text{sens})}
\end{equation}

The value of a DOR ranges from 0 to infinity, with higher values indicating better discriminatory test performance. A value of 1 means that a test does not discriminate between patients with the disorder and those without it. Values lower than 1 point to improper test interpretation (more negative tests among the diseased). The inverse of the DOR can be interpreted as the ratio of negativity odds within the diseased relative to the odds of negativity within the non-diseased. The DOR rises steeply when sensitivity or specificity becomes near perfect, as is illustrated in figure 1.\textsuperscript{6}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Behaviour of the odds ratio with changing sensitivity and specificity. Specificity: \textemdash = 0.99, \textemdash = 0.95, \textemdash = 0.80, \textemdash = 0.50}
\end{figure}
The Diagnostic Odds Ratio: A single indicator of test performance

As can be concluded from above formula's the DOR does not depend on the prevalence of the disease that the test is used for. Nevertheless, across clinical applications it is likely to depend on the spectrum of disease severity, as is the case for all other indicators of test performance. Another point to consider is that, the DOR, as a global measure, cannot be used to judge a test’s error rates, at particular prevalence’s. Two tests with an identical DOR can have very different sensitivity and specificity, with distinct clinical consequences. If a 2x2 table contains zeroes, the DOR will be undefined. Adding 0.5 to all counts in the table is a commonly used method to calculate an approximation of the DOR. Confidence intervals for range estimates and significance testing can be conventionally calculated with the following formula.

\[
\text{SE}(\log\text{DOR}) = \sqrt{\frac{1}{TP} + \frac{1}{TN} + \frac{1}{FP} + \frac{1}{FN}}
\]

A 95% confidence interval of the logDOR can then be obtained by:

\[
\log\text{DOR} \pm 1.96\text{SE}(\log\text{DOR})
\]

Calculating the antilog of this expression (back-transformation) provides the confidence interval of the DOR.

In the example of the radiological tests comparison, represented in table 3, the estimated DOR for the ultrasound in detecting peritoneal metastases is 31 ((0.69/(1-0.69))/((1-0.93)/0.93). This means that for the ultrasound the odds for positivity among subjects with peritoneal metastases is 31 times higher than the odds for positivity among subjects without peritoneal metastases. In the same way the DOR’s for the CT scan and magnetic resonance imaging (MRI) can be calculated (table 3). MR imaging has the highest DOR in detecting peritoneal metastases compared to the ultrasound and CT scan (77 versus 31 and 51 respectively). In contrast, ultrasound has the highest DOR in detecting liver metastases and lymph nodes: 54 versus 17 for CT and 15 for MRI (table 4).

If DOR’s had been presented in the original article, a quick comparison would have led to the conclusion that MR imaging performs best in diagnosing peritoneal metastases whereas ultrasound does better in diagnosing lymph node and liver metastases in this population.
Table 3 Comparison of three imaging tests in the diagnosis of peritoneal metastasis, of advanced ovarian cancer.

<table>
<thead>
<tr>
<th>Imaging test</th>
<th>Sensitivity % (95% CI)</th>
<th>Specificity % (95% CI)</th>
<th>PPV % (95% CI)</th>
<th>NPV % (95% CI)</th>
<th>LR+ (95% CI)</th>
<th>LR- (95% CI)</th>
<th>DOR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ultrasound</td>
<td>69 (58 - 80)</td>
<td>93 (90 - 97)</td>
<td>78 (68 - 89)</td>
<td>90 (85 - 94)</td>
<td>10 (6.0 - 18)</td>
<td>0.33 (23 - 5.1)</td>
<td>31 (15 - 67)</td>
</tr>
<tr>
<td>CT scan</td>
<td>92 (84 - 100)</td>
<td>81 (75 - 87)</td>
<td>61 (50 - 72)</td>
<td>97 (94 - 100)</td>
<td>5.1 (3.6 - 6.9)</td>
<td>0.10 (0.038 - 3.2)</td>
<td>51 (17 - 151)</td>
</tr>
<tr>
<td>MRI</td>
<td>95 (89 - 100)</td>
<td>80 (73 - 87)</td>
<td>59 (47 - 71)</td>
<td>98 (96 - 100)</td>
<td>4.8 (3.4 - 6.7)</td>
<td>0.06 (0.016 - 3.7)</td>
<td>77 (18 - 340)</td>
</tr>
</tbody>
</table>

Reference standard: surgical and histo-pathologic findings. Not all patients underwent all three imaging tests. Prevalence of metastasis for Ultrasound, CT scan and MRI: respectively, 68/262, 50/212 and 41/175.

Table 4 Comparison of three imaging tests in the staging of ovarian cancer; lymph node and hepatic metastasis.

<table>
<thead>
<tr>
<th>Imaging test</th>
<th>Lymph node metastases</th>
<th>Hepatic parenchymal metastasis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity % (95% CI)</td>
<td>Specificity % (95% CI)</td>
</tr>
<tr>
<td>Ultrasound</td>
<td>32 (11 - 52)</td>
<td>93 (89 - 96)</td>
</tr>
<tr>
<td>CT scan</td>
<td>43 (17 - 69)</td>
<td>89 (85 - 93)</td>
</tr>
<tr>
<td>MRI</td>
<td>38 (12 - 65)</td>
<td>84 (78 - 89)</td>
</tr>
</tbody>
</table>

Reference standard: surgical and histo-pathologic findings. Lymph node metastasis prevalence for Ultrasound, CT scan and MRI is respectively, 19/255, 14/205 and 13/171. For Hepatic parenchymal metastasis, 7/258, 5/212 and 5/165.
Polychotomous and continuous test outcomes
The performance of a test for which several cut-offs are available can be expressed by means of ROC analysis. A receiver operating characteristic (ROC) curve plots the true positive rate on the Y-axis as a function of the false positive rate on the X-axis for all possible cut-off values of the test under evaluation. The area under the curve obtained (AUC) can subsequently be calculated as an alternative single indicator of test performance.

The AUC takes values between 0 en 1, with higher values indicating better test performance. These can be interpreted as an estimate of the probability that the test correctly ranks two individuals of which one has the disease and one does not have the disease. It can alternatively be interpreted as the average sensitivity across all possible specificities.

\[
AUC = \int_0^1 \frac{1}{1 + \frac{1}{DOR} \left( \frac{x}{1 - x} \right)} dx
\]

If the DOR is constant for all possible cut-off values, the ROC curve will be symmetric (in relation to the diagonal y=x+1) and concave. In that case, a mathematical relation exists between the AUC and the DOR of a test (see formula 6). The higher the value of the DOR the higher the AUC. AUC’s of 0.85 and 0.90, for example, correspond to DOR’s of 13 and 24 respectively. An increase in AUC of 5% will almost double the DOR which is a direct consequence of scale differences: the DOR has an upper limit of infinity whereas the AUC takes values in the 0 to 1 range. For non-symmetrical ROC curves the DOR is not constant over all cut-off points. In these cases the AUC cannot be calculated from the DOR associated with a single (or a few) cut-off values.

The shape of the ROC curve and the cut-off independence of the DOR depend on the underlying distribution of test results in patients with and without the target condition. Figure 2 shows several probability densities distributions of test results in diseased and non-diseased populations. It can be observed that the DOR is reasonably constant for a large range off cut-off points on the ROC curve, but for the extremes of sensitivity and specificity the DOR rises steeply. If the original or transformed results in both diseased and non-diseased follow a logistic distribution with equal SD, the DOR is constant for all possible cut-off values (figure 2D).
The DOR in meta-analysis

The DOR offers considerable advantages in meta-analysis of diagnostic studies which combines results from different studies into summary estimates with increased precision. Meta-analysis of diagnostic tests offers statistical challenges, because of the bivariate nature of the conventional expressions of test performance. Simple pooling of sensitivity and specificity usually is inappropriate as this approach ignores threshold differences.11 18 In addition, heterogeneity may lead to an underestimation of a test's performance. The current strategy for meta-analysis, as endorsed by the Methods Working Group of the Cochrane Collaboration19, builds on the methods described by Kardaun and Kardaun and Littenberg and Moses.11 20 21 The approach by Littenberg and Moses relies on the linear regression of the logarithm of the DOR of a study (dependent variable) on an expression of the positivity threshold of that study (independent variable). If the regression line

---

Figure 2 Value of the DOR for all thresholds and distributions of test results in the non-diseased and diseased.

- A) Normal, symmetrical
- B) Lognormal, symmetrical
- C) Logistic, asymmetrical

---

The DOR in meta-analysis

The DOR offers considerable advantages in meta-analysis of diagnostic studies which combines results from different studies into summary estimates with increased precision. Meta-analysis of diagnostic tests offers statistical challenges, because of the bivariate nature of the conventional expressions of test performance. Simple pooling of sensitivity and specificity usually is inappropriate as this approach ignores threshold differences.11 18 In addition, heterogeneity may lead to an underestimation of a test's performance. The current strategy for meta-analysis, as endorsed by the Methods Working Group of the Cochrane Collaboration19, builds on the methods described by Kardaun and Kardaun and Littenberg and Moses.11 20 21 The approach by Littenberg and Moses relies on the linear regression of the logarithm of the DOR of a study (dependent variable) on an expression of the positivity threshold of that study (independent variable). If the regression line
1 The Diagnostic Odds Ratio: A single indicator of test performance

has a zero slope, the DOR is constant across studies. A summary ROC (sROC) can be produced after back transforming the regression line. The resulting sROC will be symmetric and concave. In other words, study heterogeneity can be attributed to threshold differences. In the context of the DOR, the summary DOR of the study under evaluation can be obtained from the intercept ($e^{\text{intercept}}$) of the regression line. Additional heterogeneity owing to variation in study characteristics (e.g. cohort versus case-control) or clinical characteristics (e.g. heterogeneous prior therapy) can be evaluated simultaneously by adding these variables as covariates to the regression model, leaving a corrected estimated value for the pooled DOR. The resulting parameter estimates can be (back)transformed to relative diagnostic odds ratio’s (rDOR). A rDOR of 1 indicates that the particular covariate does not affect the overall DOR. A rDOR>1 means that studies, study centres or patient subgroups with a particular characteristic have a higher DOR than studies without this characteristic. For a rDOR <1 the reverse holds. When the DOR is homogeneous across studies, the DOR’s of different studies can also be pooled directly. Homogeneity can be tested by using the Q test statistic or the H statistic.

We will illustrate the usefulness of the DOR in meta-analysis by re-analysing a meta-analysis on the diagnostic performance of two magnetic resonance angiography techniques (3D gadolinium-enhanced (3D-GD) and 2D time-of-flight (2D-TOF)) detecting peripheral arteriosclerotic occlusive disease. The separate meta-regression analysis yielded an intercept of 4.13 and a slope of 0.41 for 2D-TOF. For 3D-GD these values were respectively 5.93 and -0.37. From the intercept the summary DOR’s can be calculated, respectively $e^{4.13}=62$ and $e^{5.93}=376$. The non-zero slopes, indicated heterogeneity apart from threshold differences, which in turn limits a direct comparison of summary odds ratio’s. To explore additional variation all studies of the two techniques were put together in one regression model. Then each available covariate was examined for its effect on the diagnostic performance. Most effect had the covariates 3D-GD versus 2D-TOF technique and a covariate dealing with the post-processing technique (maximum intensity projections (MIP) in addition to transverse source images or multiplanar reformation (MIP +) versus MIP alone). Subsequently these 2 covariates were selected for the final multivariate model. The adjusted rDOR estimated was 7.5 (confidence interval: 2.8-22) for the 3D-GD and 4.5 (confidence interval: 1.5-14) for the use of MIP+. The confidence intervals did not contain the value 1. As such one can conclude that - after correction for heterogeneity - 3D-GD and the use of MIP+ have a better diagnostic performance compared to 2D-TOF and the use of MIPs alone.
The methodology of systematic reviews and meta-analysis of diagnostic tests is still evolving, with new and potentially better methods being developed, better adapted to the inherently bivariate nature of the problem. Yet the convenience of the odds ratio in statistical modelling guarantees its future role in the meta-analysis of diagnostic tests.

**Logistic regression**

The DOR offers advantages when logistic regression is used with diagnostic problems. Logistic regression can be used to construct decision rules, reflecting the combined diagnostic value of a number of diagnostic variables. Another application is the study of the added value of diagnostic tests. With a single dichotomous test the logistic regression equation reads:

\[
P(D|x) = \frac{1}{1 + \exp^{-(\alpha + \beta x)}}
\]

where \( x \) stands for the test result and the coefficients and have to be estimated. If a positive test result is coded as \( x = 1 \) and a negative as \( x = 0 \), we have

\[
P(D|\text{positive}) = \frac{1}{1 + \exp^{-(\alpha + \beta 1)}}
\]

and

\[
P(D|\text{negative}) = \frac{1}{1 + \exp^{-\alpha}}
\]

Next, one derives from expression 1,8 and 9

\[
\text{DOR} = \frac{P(D|\text{positive})}{1-P(D|\text{positive})} / \frac{P(D|\text{negative})}{1-P(D|\text{negative})} = \exp(\beta).
\]

In other words, the DOR equals the regression coefficient, after exponentiation. Logistic regression modelling has been proposed as the preferred statistical method to obtain a post-test probability of disease when results from multiple tests are available. History taking and physical examination can also be considered as individual diagnostic tests. The post test probability after having obtained test results \( x_1, x_2, \ldots, x_k \) is expressed as

\[
P(D|x_1, x_2, \ldots, x_k) = \frac{1}{1 + \exp^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}.
\]
With multiple dichotomous tests of which the results $x_1, x_2 \ldots x_k$ are coded as present (1) or absent (0), the corresponding coefficients $\beta_1, \beta_2 \ldots \beta_k$ equal the conditional logDOR. These DOR’s are conditional: they depend on the other variables that have been used in the model. If more information becomes available a new regression equation has to be constructed to obtain the proper conditional DOR. This application is illustrated by a study that aimed to assess the value of symptoms in diagnosing arrhythmias in general practice. The following equation from the logistic model was created in order to estimate the probability of arrhythmia in a patients with specific signs and symptoms:

$$P(\text{arrhythmia}) = \frac{1}{1 + \exp(-4.40 + 0.05 \times \text{age} + 0.47 \times \text{gender} + 1.11 \times \text{palpitations} + 0.78 \times \text{dyspnoea} + 0.45 \times \text{use of cardiovas.med.})}$$

Age is a continuous variable expressed in years. The odds ratio ($e^{0.05} = 1.05$) calculated from the respective coefficient does not express the diagnostic performance of the variable age, but the OR for the increase in age per year. Gender is coded 1 for males and 0 for females. The use of cardiovascular medication, palpitations and dyspnoea as recorded during consultation are coded as positive (1) or negative (0). Subsequently the conditional DOR of each dichotomous variable, adjusted for the other variables, can be estimated. For gender the DOR is $e^{0.47} = 1.6$, meaning that the odds for having arrhythmias is 1.6 times larger in males than in females. The adjusted odds ratio’s for palpitations during consultation, dyspnoea during consultation and the use of cardiovascular medication are respectively 3.0, 2.2 and 1.6.

**Discussion**

The diagnostic odds ratio as a measure of test performance combines the strengths of sensitivity and specificity, as prevalence independent indicators, with the advantage of accuracy as a single indicator. These characteristics lend the DOR particularly useful for comparing tests whenever the balance between false negative and false positive rates is not of immediate importance. These features are also highly convenient in systematic reviews and meta-analyses.

In decisions on the introduction of a test in clinical practice, we are aware that the actual balance between the true positive rate and false positive rate often matters. Whenever false positives and false negatives are weighted differentially both the prevalence and the conditional error rates of the test have to be taken into consideration to make a balanced decision. In these cases the DOR is less useful, as it does not distinguish between the two types of diagnostic mistake. If
ruling-out or ruling-in of the target condition is the primary intended use of a
test, conditional indicators such as sensitivity and specificity still have to be used.

As all available measures of test performance, the DOR of a test is unlikely to
be a test-specific constant. Its magnitude likely depends on the spectrum of disease
as well as on pre-selection through the use of other tests.\textsuperscript{6} \textsuperscript{30} Despite this universal
caveat for indicators of diagnostic tests, we feel that a more systematic use of the
odds ratio in diagnostic research can contribute to more consistent applications of
diagnostic knowledge.

Some may object that there are already too many indicators of test performance.
With such an abundance of choices, there is little need for yet another statistic.
This may be true, but it is hard to see how the selection can or should be produced.
Each of the indicators serves a different purpose. Sensitivity and specificity are
expressions of the conditional hit rates of the test. Predictive values or posterior
probabilities are the numbers that are most salient for clinical practice. The so-
called likelihood ratios come in handy for comparing the diagnostic content of
multiple possible test results and for transforming those into post-test probabilities.
Amidst those helpful indicators, the diagnostic odds ratio has a place as a single
statistic with a long history and useful statistical properties.

Acknowledgement
The authors thank Dr. Guus Hart for critically reviewing and giving useful
suggestions on earlier drafts.

References
1. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for
ovarian cancer: comparison of imaging modalities-report from the Radiological Diagnostic
The Diagnostic Odds Ratio: A single indicator of test performance


Abstract
There is ambiguity in the estimation of sensitivity and specificity in studies evaluating colonography in the screening for colon cancer. This ambiguity is due to conventional analysis methods that do not take into account multiple findings that can be obtained by colonography as well as by colonoscopy, the reference standard. In this paper the pitfalls of some frequently used methods will be discussed. In addition a new approach will be outlined that can be used to estimate sensitivity and specificity from such data. This method estimates the sensitivity and specificity, respectively, as the probability of correctly identifying true positive and true negative patients. Relative to other methods, this method takes all findings into account, including false positive and false negative lesions, as such no information is lost.
Background
When evaluating a diagnostic test one is interested in how well the test is able to
distinguish patients with the target condition from those without that condition.
Different measures of test performance are available to express a test’s performance.1
Well known properties are the sensitivity and specificity of a test. These two
measures indicate the proportion of patients correctly classified by the index test
in the population of diseased and non-diseased persons, respectively, as diagnosed
by the reference standard.

In addition to making a distinction between diseased and non-diseased
persons, tests can also be used to document the size, severity, location or number
of target lesions. An example can be found in the evaluation of CT or MRI (2D
or 3D) colonography (virtual colonography). Each technique can result in the
detection of multiple polyps. In the analysis of such studies, lesions detected on
colonography are compared with the findings of conventional colonoscopy. Such
comparisons may or may not result in a match, whenever lesions are found on
both tests.

It is not clear how the number of matches and mismatches can be transformed
into one of the more conventional test statistics, such as sensitivity and specificity.
By discussing various methods used in the literature to analyse data from such
colonography evaluations, we will show that the estimation of sensitivity and
specificity is not without pitfalls. In addition we present a new method that can
be applied to this kind of data. The application of these methods will be illustrated
with data from a recently performed study to evaluate the test characteristics of
CT colonography (CTC) at different levels of radiation dose. In this study 50
patients at risk for colorectal cancer were evaluated for the presence of one or
more polyps (or other findings e.g neoplasms). Colonoscopy was used as the
reference standard. Herein only the data at mAs levels of 100 will be used and
lesions of all sizes will be considered. Detailed information can be found elsewhere.2

Methods

Per-polyp analysis
In many studies of CTC a per-polyp based approach is used in the evaluation, in
which the polyps are the unit of analysis.3-7 In these studies, the researchers compare
polyps found on conventional colonoscopy with the findings of CTC, estimating
the per-polyp sensitivity of the latter.

In our study, for example, colonoscopy revealed 62 polyps found in 27
patients. Of these polyps 23 could also be detected by CTC. The sensitivity is
then estimated as 23/62 or 37% (table 1), indicating a probability of 0.37 that a polyp detected by colonoscopy can also be detected by CTC. Alternatively, one can start from the polyps found on CTC and see whether these can be verified by conventional colonoscopy, this way calculating the per-polyp positive predictive value.

One must consider that the estimated sensitivity can be biased and its statistical precision overstated because of clustering of polyps within patients. Such clustering can lead to correlation of diagnostic on multiple polyps in the same patient. In addition, individual patients with multiple polyps would effectively carry more weight in the calculation of sensitivity relative to patients with only few polyps. A correction for clustering of polyps within patients is needed to guarantee that k polyps in a single patient have less weight in the estimation of sensitivity than a series of single polyps in k patients.8 9 The per-polyp specificity of the test is a lot harder to define. In principle there is an infinite number of negative results: each potential location of a polyp in the colon. As a consequence the per-polyp specificity and negative predictive value are undefined.

Table 1 The sensitivity and specificity as estimated by different methods. Data regarding polyps of all sizes.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Uncorrected</td>
<td>Corrected</td>
</tr>
<tr>
<td>Per-polyp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unmatched</td>
<td>93 (83 to 100)</td>
<td>-</td>
</tr>
<tr>
<td>FFP Conditional on index test</td>
<td>89 (74 to 100)</td>
<td>-</td>
</tr>
<tr>
<td>FFP Conditional on reference test</td>
<td>59 (41 to 78)</td>
<td>-</td>
</tr>
<tr>
<td>Combined per-polyp/per-patient</td>
<td>37 (25 to 49)</td>
<td>37 (25 to 49)</td>
</tr>
<tr>
<td>FROC approach</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per-polyp</td>
<td>49 (11 to 86)</td>
<td>-</td>
</tr>
<tr>
<td>Per-patient</td>
<td>79 (24 to 99)</td>
<td>26 (19 to 36)</td>
</tr>
</tbody>
</table>

* In the corrected results, clustering of polyps within patients is taken into account.

Per-Patient analysis
A per-patient analysis is often done as an alternative. In this approach the patient is the unit of analysis.510-14 A patient is classified as positive by the reference test if at least one polyp of a pre-specified size is present and negative if no such lesion is detected. A true positive patient is a patient with at least one polyp detected by the reference standard and at least one lesion on the index test, regardless its location. In our study 27 of the 50 patients were found to have at least one polyp...
The sensitivity and specificity of tests that locate multiple lesions; evaluating colonography

by colonoscopy. Of these patients 25 also had at least one lesion at CTC, resulting in a per-patient sensitivity of 93% (25/27). Of the remaining 23 patients without lesions detected at colonoscopy 6 were also negative on CTC, giving a specificity of 26% (6/23).

This analytic approach is relevant in the context of clinical practice. CTC is mainly propagated as a screening tool and, if any lesion is found, that patient is subsequently scheduled for conventional colonoscopy, a technique that also allows removal of the polyps.\textsuperscript{4,10} The disadvantage of this approach is that all patients with a lesion on CTC and a polyp found by colonoscopy are counted as true positives. It is possible that these two lesions are found at different sites in the colon. For convenience we call this 5\textsuperscript{th} category of patients, the ‘floating false positives’ (FFP).

Counting these FFP’s as true positive patients may seem arbitrary, as they reflect a lack of correspondence between findings on the index test and those of the reference standard. If within the study design lesions have not been matched, the FFP’s are inevitably regarded as true positive patients. If lesions have been matched, based on size and location, there are three cells within the 2x2 table possible for the FFP’s. In our study there were nine FFP patients, thus nine patients with on both test at least one lesions, but where none of these lesions matched. One possibility is regarding the lesions as unmatched, hence interpreting the FFP’s as discussed above.\textsuperscript{3,4,15} Alternatively, if one conditions on the results of the index test, that is by fixing the row totals, the FFP’s are considered as false positives (see table 2).

<table>
<thead>
<tr>
<th>Reference standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index test</td>
</tr>
<tr>
<td>pos</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17+9</td>
</tr>
<tr>
<td>42</td>
</tr>
<tr>
<td>neg</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>32</td>
</tr>
<tr>
<td>50</td>
</tr>
</tbody>
</table>

Estimated sensitivity = 16/18 = 0.89, specificity = 6/32 = 0.19.

If one conditions on the reference standard results (fixing the column totals), the FFP’s are considered as false negatives (see table 3).
The cell in which the FFP patients are put clearly affects the sensitivity and specificity (see table 1). The larger the proportion of FFP's, the larger the difference in the value of these accuracy estimates.

An illustration of the ambiguity that can follow from these FFP results can be found in the report of a study performed by Fenlon and colleagues. In that study CTC and conventional colonoscopy were compared in 100 patients of whom 49 had polyps. The following results can be found in the paper: 42 true positives, 41 true negatives, 8 false positives and 9 false negatives. Adding up the these totals gives us 51 patients with polyps, not 49 as given in the beginning of the results section of that paper. This discrepancy is probably due to 2 FFP patients. Since the authors did not specify whether this type of positives were present and, if so, to which cell they were assigned, the way in which the results are presented impedes a clear interpretation of the test’s properties.

Per-Segment analysis
The disadvantage of the per-patient approach is that the multiple findings, and their location in a single patient are ignored, resulting in a loss of information in the analysis. Patients with multiple false positive results are treated in the same way as patients with a single false positive result without affecting the estimated specificity of the test.

Some authors propose to divide the organ of interest in a number of segments, - the ascending colon for example - and take the segments as the unit of analysis. This approach has been used, for example in studies of the performance of diagnostic modalities in staging colorectal cancer, detecting liver metastases, staging prostate cancer, detecting musculoskeletal cancer, and MR Angiography. In each patient, every segment is scored for the existence of polyps. A true positive segment is a segment with at least one positive lesion both on colonoscopy and on
CTC. A true negative segment is one in which nothing is seen on both tests. The sensitivity can then be estimated as the number of positive segments as found on CTC relative to with number of segments with at least one polyp on colonoscopy. The specificity is estimated as the proportion of negative segments found by colonoscopy that were also negative on CTC.

The problem of FFP has not disappeared in this approach. The proportion of FFP segments now depends on the number of segments, with the probability of FFP decreasing with an increasing number of segments. Increasing the number of segments will increase the number of true negative segments and this will inflate the specificity. In addition cluster dependency will exist between segments from the same patient and has to be corrected for. The design of our study did not include the recording of segment specific information, so we cannot report on this kind of analysis.

Combined per-polyp/per-patient analysis
To bypass the impossibility of estimating a per-polyp specificity some authors have turned to a mixed approach: presenting a per-polyp sensitivity and a per-patient specificity.\(^7\)\(^2\)\(^0\) If specificity is calculated as the proportion of patients without lesions in the group of patients who test negative for polyps on colonoscopy then the issue of FFP patients is avoided. FFP’s affect the sensitivity in a per-patient approach. Although definition and interpretation of the single sensitivity and specificity are straightforward in this mixed approach, one has to be aware that this strategy neglects a key aspect of diagnostic test evaluation: the close relation and implicit threshold balance between sensitivity and specificity. Likelihood ratio’s, well known in the Bayesian updating of prior probabilities, cannot be derived from these sensitivity and specificity values, where one is based on a per-patient analysis and the other is polyp based.

FROC methodology
About fifteen years ago Chakraborty introduced a method within the field of radiology to estimate free-response receiver operating characteristics (FROC) curves for tests that are able to locate multiple findings.\(^2\)\(^1\)\(^\text{-}\)\(^2\)\(^3\) The FROC method was originally developed for observer performance studies of imaging tests where the observer classifies each image with one of several (usually five) confidence ratings. These ratings reflect the observer’s confidence that the image is abnormal.\(^2\)\(^1\)\(^-\)\(^2\)\(^5\) This FROC method can be adapted to estimate sensitivity and specificity of tests that can locate multiple lesions whenever dichotomous (positive or negative) outcomes are preferred.
The FROC approach estimates a false positive rate, from which a specificity can be calculated. This approach assumes that the number of false positives follows a Poisson distribution, in other words, the probability of a specific number of false positives in the i-th patient is given by the equation:

$$Pr(X_i = x_i) = \frac{e^{-\mu} \mu^{x_i}}{x_i!}$$

where the parameter $\mu$ represents the expected number of FP lesions per patient. When there are $n$ patients, the rate parameter $\mu$ is estimated as

$$\frac{1}{n} \sum x_i$$

The specificity is then defined as the probability that no false positives are observed in a patient $Pr(X_i=0)$. The probability that this occurs is equal to $Pr(X=0)=e^{-\mu}$.

The assumption of a common FP rate across patients is somewhat too restrictive, and a common method to relax it is to assume that the rate parameter varies between patients, in which case:

$$Pr(X_i = x_i) = \frac{e^{-\mu_i} \mu_i^{x_i}}{x_i!}$$

The variation in the FP rates $\mu_i$ between patients can be described by a distribution $g(\mu)$. In this way the mean specificity is defined as $Pr(X=0)=\int e^{-\mu} g(\mu) d\mu$. This integral can be interpreted as a weighted average of the specificity values across patients, with the weights given by the distribution function $g(\mu)$. This function is unknown but can be estimated from the data. In order to do so, choices must be made about the functional form of $g(\mu)$. Most common choices for $g(\mu)$ are the gamma distribution, but we prefer a fully non-parametric distribution. The latter approach takes cluster variation of polyps between patients into account and is preferable in the presence of variability. The approach just presented allows the estimation of a per-patient specificity using all information available. As it is based on a joint analysis of all patients, those with and without proven polyps, the estimated per patient specificity is automatically corrected for FFP patients.

A somewhat similar approach can be used to model the sensitivity. Assume that the number of true positive polyps $Y_i$ in patient $i$, having $k_i$ polyps as defined by colonoscopy follows a binomial distribution with specificity $P_i$. If the between-patient variation in $P_i$ is described by a nonparametric distribution $g(p)$, the probability of a true positive polyp in a randomly selected patient - the per-polyp sensitivity - can be expressed as a similar weighted average $\int g(p)dp$.
Notice that the per-patient sensitivity can be written as a function of the per-polyp sensitivity ($\phi$) and the mean number of polyps ($K$) per patient which approximately follows the binomial distribution: $1-(1-\phi)^K$. In our data $K=62/27=2.3$ and $\phi=0.49$. The per-patient sensitivity can then be estimated as 0.79 (see table 1).

A disadvantage of this approach is that it analyzes sensitivity and specificity separately, without accounting for the tradeoff between the two quantities. The computations are somewhat complex, though available through routines in packages like SAS and Egret. A computer program that simultaneously estimates the specificity and sensitivities can be obtained free of charge from the first author.

Discussion
We have discussed various methods to express the results of evaluation studies of colonographic techniques. These methods can produce widely diverging estimates of test performance, as can be inferred from table 1. The variety of methods can be attributed to the strong inclination of investigators to force the results of such studies in a 2x2 table and to calculate sensitivity and specificity. With tests that are able to detect and locate multiple lesions, the binary patient-based nature of these measures is lost and the classical 2x2 table loses its straightforward appeal. Only with additional assumptions and simplifications can one arrive at a 2x2 table. However, the resulting sensitivity and specificity may no longer express a test’s true diagnostic performance.

In this paper, we have introduced a convenient approach that is able to incorporate all lesions detected in the analysis. As such our approach bypasses ambivalent FFP patients. If a random effects model is considered it directly takes into account dependency between lesions found in the same patient. However, the modified FROC approach analyzes sensitivity and specificity separately, without accounting for the tradeoff between the two quantities.

As far as we know, FFP patients have never been reported in the articles that used a per-patient based approach to estimate sensitivity and specificity. This omission might have led to ambiguity of results, and a hampered critical appraisal by interested readers. In addition it raises the question whether the today’s impression of the diagnostic performance of virtual colonography is correct. As a promising technique, virtual colonography deserves a proper evaluation. This can only happen if investigators look critical at the methodology used and turn to clear and complete reporting of their results.
References
2 The sensitivity and specificity of tests that locate multiple lesions; evaluating colonography


3 Direct pooling of sensitivity and specificity using bivariate models in meta-analysis of studies of diagnostic accuracy

Johannes B. Reitsma, Afina S. Glas, Anne W.S. Rutjes, Rob J.P.M. Scholten, Patrick M.M. Bossuyt, Aeilko H. Zwinderman

Abstract
Diagnostic accuracy studies most often report estimates of sensitivity and specificity to describe the properties of a diagnostic test. Separate pooling of sensitivity and specificity is usually inappropriate because of the negative correlation between these two measures due to differences in the way studies defined positive and negative test results (threshold differences).

The summary Receiver Operating Characteristics (ROC) approach has become the method of choice for pooling pairs of sensitivity and specificity. This method transforms sensitivity and specificity into a single measure of accuracy, the diagnostic odds ratio, to deal with the negative correlation between sensitivity and specificity. The disadvantage of using the diagnostic odds ratio as the outcome measure is that summary estimates of sensitivity and specificity are not directly available. Furthermore, by removing the effect of a possible threshold when comparing different types of test, relevant clinical differences in test performance may get lost.

We advocate the use of bivariate models to preserve the two-dimensional nature of the underlying data, directly leading to pooled estimates of sensitivity and specificity, and at the same time incorporate any correlation that might exist. Explanatory variables can be added to the model and lead to separate effects of a particular variable on sensitivity and specificity, rather than a net effect on the odds ratio scale as in the summary ROC approach. The statistical properties of the model are sound and flexible, leading to a more simplified approach to the meta-analysis of diagnostic data.
Introduction
Diagnostic accuracy studies are a vital step in the evaluation of new diagnostic technologies. In diagnostic accuracy studies the results of the test under evaluation are compared with the best available method, the reference standard. Accuracy refers to the amount of agreement between the information from the test under evaluation and the reference standard. There are several different measures of diagnostic accuracy, but the majority of diagnostic accuracy studies present estimates of sensitivity and specificity, either alone or in combination with other measures.

Combining evidence from several studies enables researchers to produce a summary measure with less statistical uncertainty and provides the opportunity to examine the effect of differences in clinical and design related factors between studies. These potential advantages have led to a sharp increase in the number of meta-analyses during the 1990s.

Because the majority of diagnostic papers report estimates of sensitivity and specificity, meta-analytic approaches have focused on these measures. Pooling pairs of sensitivity and specificity is not straightforward, because these measures are often negatively correlated across studies. Separate pooling of these measures is then inappropriate.

The summary Receiver Operating Characteristic (ROC) approach has become the method of choice for pooling pairs of sensitivity and specificity. The summary ROC approach converts pairs of sensitivity and specificity into a single measure, the diagnostic odds ratio, to deal with the correlation between sensitivity and specificity. The disadvantage of a single measure of diagnostic accuracy is that it does not distinguish between the ability of detecting the sick (sensitivity) and identifying the well (specificity). Discriminating between these abilities is important to determine the optimal use of a test in clinical practice. The bivariate model we propose has the distinct advantage of preserving the two-dimensional nature of the underlying data. It directly leads to pooled estimates of sensitivity and specificity, acknowledging any possible (negative) correlation between these two measures. We will discuss both approaches and illustrate their use by reanalyzing the data from a published meta-analysis.

Pooling pairs of sensitivity and specificity: why simple methods fail
The starting point is a set of individual studies that have all presented estimates of sensitivity and specificity. One intuitive approach to obtain a summary estimate would be to pool sensitivities and specificities separately using standard methods to combine proportions. However, sensitivity and specificity are often negatively
correlated across studies, and ignoring this correlation would be inappropriate.6 10 11

An important cause for this negative correlation between sensitivity and specificity is that studies may have used different thresholds to define positive and negative test results. In some cases this may have been done explicitly, for example studies that used different cut-off points to classify a continuous biochemical measurement as positive or negative. In other situations there may have been naturally occurring variations in thresholds between studies due to differences in observers, laboratories, or machines. Unlike other sources of variation, a difference in threshold leads to a particular pattern between sensitivity and specificity. This pattern is well known from studies showing the effect of different cutoffs in case of a biochemical test with a continuous outcome.19-21 Lowering the cutoff value will then lead to more patients with positive results, thereby increasing the number of true positives but also the number of false results. This means that sensitivity will be higher but specificity lower. This trade off between sensitivity and specificity leads to a concave, shoulder-like curve when sensitivity is plotted against 1 minus specificity, the receiver operating characteristic (ROC) curve. In many publications involving ROC plots, sensitivity is referred to as the true positive rate (TPR) and 1-minus specificity as the false positive rate (FPR).

In figure 1 we plotted the sensitivity and specificity from 17 studies examining the diagnostic accuracy of computed tomography for the diagnosis of lymph node metastasis in women with cervical cancer.22 It is a nice example of the three features that many meta-analyses have in common: (1) large variation in both sensitivity and specificity; (2) negative correlation between sensitivity and specificity across studies; (3) substantial differences in the size of the individual studies. In the next paragraphs, we discuss how the summary ROC and the bivariate approach deal with this situation.
The summary ROC approach
We provide a short description of the summary ROC approach as outlined by Moses and Littenberg (more details can be found in these publications).

To illustrate the use of the summary ROC approach, we re-analyzed the data of a published meta-analysis. In this meta-analysis three imaging techniques were compared for the diagnosis of lymph node metastasis in women with cervical cancer. Forty-four studies in total were included; 17 studies evaluated lymphangiography, another 17 studies examined computed tomography and the remaining 10 studies focused on magnetic resonance imaging. Diagnosis of metastatic disease by lymphangiography (LAG) relies on the presence of nodal filling defects, whereas computed tomography (CT) and magnetic resonance imaging (MRI) rely on nodal enlargement.

The summary ROC approach starts with plotting the observed pairs of sensitivity and specificity of each study in the ROC space (see figure 2). The aim of the summary ROC approach is to find a smooth curve through these points. The key step is to transform the TPR and FPR scale of the ROC graph in such a way that the relation becomes more linear and a straight line can be fitted through the data.
The following transformations of TPR and FPR are used. D is defined as the difference in the logit transformed values of TPR and FPR, while S is the sum of these same logits. In equation:

\[ D = \ln \left( \frac{TPR}{1 - TPR} \right) - \ln \left( \frac{FPR}{1 - FPR} \right) = \ln(DOR) \]

and

\[ S = \ln \left( \frac{TPR}{1 - TPR} \right) + \ln \left( \frac{FPR}{1 - FPR} \right) \]

D is equivalent to the (log of the) diagnostic odds ratio (DOR), a single overall indicator of diagnostic accuracy. The DOR shows how more often a positive test results occurs among patients with the condition of interest compared to those patients without the condition. S relates to the test threshold: it has a value of 0 when sensitivity equals specificity, it is positive in studies where sensitivity is higher than specificity and negative when specificity is higher.

The linear regression line that is fitted through the transformed points of the ROC graph is given by the following equation:

\[ D = \alpha + \beta \cdot S \]

Similar to the original analysis we fitted three different regression lines, one for each imaging modality. The intercepts (\( \alpha \)) and the slopes (\( \beta \)) of these three lines are given in table 1. In the final step, the regression line(s) of equation 3 is
3 Direct pooling of sensitivity and specificity using bivariate models in meta-analysis of studies of diagnostic accuracy

transformed back to the original axes of the ROC to obtain, the summary ROC curve. Panel A of figure 3 shows the three summary ROC curves, one for each imaging modality.

Table 1. Intercepts and slopes of the linear regression line of underlying the summary ROC approach. Separate lines are fitted for each of the three imaging modalities. Comparison of accuracy using Q-points and the diagnostic odds ratio at the average value of the threshold.

<table>
<thead>
<tr>
<th>Imaging modality</th>
<th>Parameter</th>
<th>DOR at mean of S (95% CI)</th>
<th>Q (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept α (se)</td>
<td>Coefficient for S β (se)</td>
<td></td>
</tr>
<tr>
<td>LAG</td>
<td>2.25 (0.32)</td>
<td>-0.35 (0.19)</td>
<td>19.59 (9.85 to 38.98)</td>
</tr>
<tr>
<td>CT</td>
<td>3.11 (0.49)</td>
<td>0.25 (0.16)</td>
<td>13.31 (7.52 to 23.53)</td>
</tr>
<tr>
<td>MRI</td>
<td>3.80 (0.61)</td>
<td>0.28 (0.18)</td>
<td>25.47 (12.25 to 52.96)</td>
</tr>
<tr>
<td>P-value LAG vs. CT</td>
<td>0.40</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>P-value LAG vs. MRI</td>
<td>0.61</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>P-value CT vs. MRI</td>
<td>0.17</td>
<td>0.37</td>
<td></td>
</tr>
</tbody>
</table>

LAG = lymphangiography; CT = computed tomography; MRI = magnetic resonance imaging; Q-point = point on summary ROC where sensitivity equals specificity; DOR = diagnostic odds ratio.

Interpretation and results

The interpretation of intercept and the slope of the linear regression model is not straightforward. When the diagnostic odds ratio (DOR) does not change linearly with S (e.g., β ≈ 0), the intercept would provide a summary estimate of the DOR. When the DOR does vary with S, the coefficient of the slope (β) has no direct interpretation, but has a considerable effect on the shape of the summary ROC curve.¹⁵

The disadvantage of the diagnostic odds ratio as the outcome parameter is that summary estimates of sensitivity and specificity are not directly available. It is only possible to obtain an estimate of sensitivity by specifying a value of specificity, or vice-versa. Most meta-analyses have used the Q-point, which is the point on the summary ROC curve where sensitivity equals specificity. This Q-point is located where the diagonal line running from the top left corner to the lower right corner intersects the summary ROC curve (see panel B figure 2). Unfortunately, the Q-point may lead to summary values of sensitivity and specificity that are near or even outside the range of values from the original studies (see Q-point of MR+ in panel A of figure 3).

Q-points have also been used to test for a difference in overall accuracy between diagnostic tests. The rationale is that Q-points remove the effect of possible difference in threshold by comparing the diagnostic odds ratios at a specific value of S, namely zero. However, testing at a different value of S could lead to different conclusions if the diagnostic odds ratio of one or both tests varies with S.
example there is a statistically significant difference in accuracy at the Q-point between lymphangiography and MRI, but at the overall mean value of S this difference in diagnostic odds ratio no longer exist (see table 2 and figure 3).

Table 2. Summary estimates for sensitivity, specificity and diagnostic odds ratio from the bivariate model. Comparison between three imaging modalities.

<table>
<thead>
<tr>
<th>Imaging modality</th>
<th>Mean sensitivity (95% CI)</th>
<th>Mean specificity (95% CI)</th>
<th>Mean DOR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAG</td>
<td>0.67 (0.57 to 0.76)</td>
<td>0.80 (0.73 to 0.85)</td>
<td>8.13 (5.16 to 12.82)</td>
</tr>
<tr>
<td>CT</td>
<td>0.49 (0.37 to 0.61)</td>
<td>0.92 (0.88 to 0.95)</td>
<td>11.34 (6.66 to 19.30)</td>
</tr>
<tr>
<td>MRI</td>
<td>0.56 (0.41 to 0.70)</td>
<td>0.94 (0.90 to 0.97)</td>
<td>21.42 (10.81 to 42.45)</td>
</tr>
<tr>
<td>P-value LAG vs. CT</td>
<td>0.023</td>
<td>0.0002</td>
<td>0.35</td>
</tr>
<tr>
<td>P-value LAG vs. MRI</td>
<td>0.23</td>
<td>0.0001</td>
<td>0.021</td>
</tr>
<tr>
<td>P-value CT vs. MRI</td>
<td>0.47</td>
<td>0.34</td>
<td>0.15</td>
</tr>
</tbody>
</table>

LAG = lymphangiography; CT = computed tomography; MRI = magnetic resonance imaging; DOR = diagnostic odds ratio.

The bivariate model

The bivariate model uses a different starting point for pooling estimates of sensitivity and specificity from a group of individual studies. Rather than transforming these two effect measures into a single indicator of diagnostic accuracy as in the summary ROC approach, the bivariate model preserves the two-dimensional nature of the data throughout the analysis.
3 Direct pooling of sensitivity and specificity using bivariate models in meta-analysis of studies of diagnostic accuracy

The bivariate model is based on the following line of reasoning. We assume that the set of sensitivities from individual studies (after logit transformation) are approximately normally distributed around a mean value with a certain amount of variability around this mean. The same is true for the specificities of these studies. The problem of the possible (negative) correlation between sensitivity and specificity across studies is addressed by explicitly incorporating this correlation into the analysis. Combining two normal distributions, in this case the logit transformed sensitivities and specificities, while acknowledging the possible covariance (correlation) between these two measures, leads to the bivariate normal distribution.

We extended the bivariate model by incorporating the precision by which sensitivity and specificity have been measured in each study using the approach of Van Houwelingen. The central issue in the bivariate analysis of sensitivity and specificity is similar to the problem of analyzing repeated outcomes in longitudinal studies. In longitudinal studies we also expect correlation between measurements from the same individual at different points in time, and take this into account in the statistical analysis. The bivariate diagnostic model can be viewed as a longitudinal analysis with measurements at two time points (corresponding to sensitivity and specificity) within individuals (corresponding to individual studies). A more technical description can be found in Appendix I.

The logit-transformed sensitivity and specificity of each study are the two outcome parameters in the bivariate model. In a single model we will estimate the mean value of logit sensitivity and specificity and their corresponding variances, but also the covariance between logit sensitivity and specificity across studies. These bivariate models can be analyzed using linear mixed model techniques that are now widely available in statistical packages. Annotated commands to run the bivariate model using the linear mixed model procedure in SAS are given in Appendix II. The bivariate model will produce the following parameters of interest:

- The mean value of logit sensitivity with a 95% confidence interval (CI), which can be easily transformed to the mean sensitivity and 95% CI on the original scale.
- The mean value of logit specificity with a 95% CI.
- The correlation between logit sensitivity and specificity.
- The mean value of the log of the diagnostic odds ratio with a 95% CI.

We will now use the bivariate model to re-analyze the same dataset as we used in the summary ROC approach.
Interpretation of the results of the bivariate model
The bivariate model directly provides summary estimates of (logit) sensitivity and specificity with corresponding 95% CI for the three imaging modalities (see table 2). Because of the bivariate nature of the analysis we can also test for differences in either sensitivity or specificity or both between the three modalities.

These results show that the mean value of both sensitivity and specificity of LAG are significantly different from that of CT and MRI. On average, LAG has a higher sensitivity but a lower specificity. There are no statistically significant differences in mean sensitivity or specificity between CT and MRI.

The difference between LAG and CT/MRI could be viewed as the result of a threshold effect. It shows that LAG is a more sensitive test, but at the expense of more false positive test results, and hence a lower specificity. It means that the overall accuracy after ‘correction’ for threshold differences (e.g., diagnostic odds ratio, Q-point) is comparable between the three techniques. This explains the results of the summary ROC approach, no differences among the three techniques. The results of the bivariate model show a more complete picture: a difference in sensitivity and specificity between LAG and the other two techniques, but no difference in overall accuracy.

The visual power of the ROC plot can still be used to visualize the results of the bivariate model. The 95 per cent coverage region of the estimated bivariate distribution of logit sensitivity and specificity can be transformed back to the original ROC axes. This ellipse around the mean sensitivity and specificity of each modality will show the region that contains likely combinations of summary values of sensitivity and specificity. These ellipses nicely show the differences in sensitivity and specificity of LAG compared to CT and MRI (see panel B of figure 3).

Comparison between the summary ROC and the bivariate approach
Ideally, the statistical method for the meta-analysis of diagnostic data should meet the following requirements. First, the approach uses a summary measure that is clinically meaningful. Second, the approach is flexible so that variables can be added to the model to examine differences between tests and to explore possible sources of heterogeneity. Third, the underlying statistical properties are sound. Fourth, the approach is easy to use and avoids complex programming or calculations. We will compare both approaches and focus on these key issues.

Summary measure of interest
The majority of diagnostic accuracy studies report their results in pairs of sensitivity and specificity. The bivariate model directly provides summary estimates for these
well-known measures of diagnostic accuracy. The 95% confidence ellipse around the mean value of sensitivity and specificity provides a nice summary of the results, because it takes into account the correlation between sensitivity and specificity, and the amount of between-study variation (random effects model).

The summary ROC approach focuses on the diagnostic odds ratio, and therefore it does not yield a unique summary estimate of sensitivity and specificity. This greatly limits its clinical application. Summary values of sensitivity and specificity belonging to the Q-point are just an arbitrary choice of possible values, and can be highly misleading (compare the values of the Q-point of MRI with the pooled estimates of sensitivity and specificity from the bivariate model).

Comparing accuracy between tests and exploring sources of heterogeneity
The summary ROC approach uses the odds ratio to compare diagnostic accuracy between tests. This removes the effect of a possible difference in threshold, but at the same time it can remove important clinical differences in test performance. In this example, LAG was more sensitive and less specific compared to the other two techniques, which might indicate that it can be an important modality for ruling out lymphadenopathy. In the bivariate model we can specifically test whether there is a difference in sensitivity, specificity, or both.

Examining, quantifying and explaining sources of heterogeneity in meta-analysis is a major issue, in particular for diagnostic studies as there are many possible differences in design, in the selection of patients, and in test protocol between studies. Again, the effect of these differences in design and conduct will be estimated as changes on the diagnostic odds ratio scale. An unchanged odds ratio however, may obscure the effect of a characteristic that increases sensitivity but at the same time lowers specificity, or vice-versa. The bivariate model will show separate effects on sensitivity and specificity. A net effect on the odds ratio scale is still available.

Sound statistical framework
The statistical properties of the bivariate approach are sound as it deals with all the requirements of advanced meta-analysis. The bivariate model takes into account the differences in precision by which sensitivity and specificity have been measured across studies, and it correctly incorporates and estimates between-study variation (random effects model). The flexibility and soundness of the bivariate model become evident when we consider three special situations.

The first situation is where the variation in sensitivity and specificity across studies can be adequately explained by a single diagnostic odds ratio. In other
words, changes in (logit) sensitivity are accompanied by similar, but opposite changes in specificity resulting in a constant diagnostic odds ratio, except from some random error. In this case the bivariate model will produce the same diagnostic odds ratio and 95% confidence interval as a direct pooling of the odds ratios using the well-established random effects approach of DerSimonian and Laird.32

In the opposite situation, when there is no correlation between sensitivity and specificity across studies, the results of the bivariate model will be similar to a separate random effects pooling of (logit) sensitivity and specificity. Because the bivariate integrates and estimates the correlation between sensitivity and specificity, it will produce correct results whether this correlation is high, low or medium.

Another example of the flexibility of the bivariate model is in situations where the variation in one measure between studies is small, but there is substantial variation in specificity. We could then proceed with a model that obtains a fixed effect summary measure of sensitivity and a random effects pooling of specificity. To fit this model we would only have to drop the variable indicating sensitivity from the random statement of the proc mixed model (see appendix I).

The summary ROC approach based on the linear regression of D on S has two statistical shortcomings. First, there is measurement error in both the dependent (D) and the independent variable (S). There are well-known methods that take into account measurement error in both the dependent and independent variables, but these are seldom used in the summary ROC-approach. However, a more serious problem is ignoring the covariance that exists between D and S, because they are defined as the difference and sum of the same two measures.

The problems in the interpretation of the results and the statistical shortcomings of the summary ROC approach have been described before.10 12 33 Gatsonis et al. have developed a technique which they called the hierarchical summary ROC approach.10 34 Their aim was also to obtain meaningful summary estimates of sensitivity and specificity, and to improve the handling of within- and between-study variation. Because we used the same meta-analysis as they reanalyzed, we can directly compare their results with ours. Although their starting point is different, the summary estimates and 95% CI of sensitivity and specificity of both approaches are almost identical. However, their approach comes at a price: it is based on Markov Chain Monte Carlo simulations. This technique requires programming, simulation, evaluation of model convergence and adequacy, and synthesis of simulation results. This brings us to the fourth issue: easiness of use.
Easy to use
The development of mixed model technology in commercial software means that bivariate models can now be analysed with standard statistical packages. We have shown in appendix I how these models can be set up using standard procedures with SAS software without the need for complex calculations.

The hallmark of the bivariate model of incorporating the amount of correlation between sensitivity and specificity simplifies the overall approach to meta-analysis as outlined in recent guidelines. Centrally in their flow charts is the examination of the degree of correlation between sensitivity and specificity. If moderate correlation is present, the summary ROC approach is advocated and if correlation is small, separate pooling of sensitivity and specificity is promoted. Bivariate models will automatically deal with all these situations.

Diagnostic accuracy studies: quality of methods and reporting
Meta-analysis of diagnostic studies will remain difficult despite the availability of flexible and advanced statistical models to analyze such data for three main reasons. First, a systematic review of diagnostic accuracy studies, like any other review, is threatened by publication bias. Second, many reports of studies of diagnostic accuracy lack information on key elements of design and conduct. Without complete and accurate reporting we cannot correctly classify a study according to certain features of design or conduct. This hampers the possibility to investigate the potential for bias generated by these features. Third, many studies on diagnostic accuracy have major shortcomings in design or conduct. Health care workers need evidence from well-designed studies to make informed choices. Summarization remains useless in the absence of set of results from studies of premium quality. In all, there is a strong need to improve the methodological quality of diagnostic studies in addition to better standards of reporting.

References
3 Direct pooling of sensitivity and specificity using bivariate models in meta-analysis of studies of diagnostic accuracy

Appendix I. Technical description of the bivariate model

We consider individual studies that have reported sensitivity \( p_{A,i} \) determined in \( n_A \) individuals with the condition of interest and specificity \( p_{B,i} \) measured in \( n_B \) subjects without the condition. We define \( \theta_{A,i} \) as the logit-transformed sensitivity in study \( i \) and \( \theta_{B,i} \) as the logit-transformed specificity. When \( n_A \) and \( n_B \) are large and \( 0 < p_{A,i}; p_{B,i} < 1 \) then the estimated logit transformed sensitivity and specificity are normally distributed with corresponding variances given by

\[
\begin{align*}
\sigma^2_{A,i} &= \frac{1}{n_A \cdot p_{A,i} \cdot (1 - p_{A,i})} \\
\sigma^2_{B,i} &= \frac{1}{n_B \cdot p_{B,i} \cdot (1 - p_{B,i})}
\end{align*}
\]

We assume that the true outcome measures of an individual study (pair of \( \theta_{A,i} \) and \( \theta_{B,i} \)) follow a bivariate normal distribution around some common mean value of logit sensitivity and specificity

\[
\begin{pmatrix}
\theta_{A,i} \\
\theta_{B,i}
\end{pmatrix} \sim N
\begin{pmatrix}
\theta_A \\
\theta_B
\end{pmatrix}
\Sigma
\]

\[
\Sigma = \begin{pmatrix}
\sigma_A^2 & \sigma_{AB} \\
\sigma_{AB} & \sigma_B^2
\end{pmatrix}
\]

\( \sigma_A \) and \( \sigma_B \) describe the variability among studies in logit sensitivity and specificity. \( \sigma_{AB} \) is the covariance between logit sensitivity and specificity. The stochastic nature of sensitivity and specificity can easily be incorporated into the bivariate model if we treat the observed variance of logit sensitivity and specificity as fixed quantities, the standard approach in meta-analysis. The resulting marginal model then becomes

\[
\begin{pmatrix}
\theta_{A,i} \\
\theta_{B,i}
\end{pmatrix} \sim N
\begin{pmatrix}
\theta_A \\
\theta_B
\end{pmatrix}
\Sigma + C_i
\]

with \( C_i \) being a diagonal matrix containing the \( S_i \)’s.

This model can be fitted by Maximum Likelihood methods based on the EM algorithm. The SAS proc mixed module is a convenient mixed model programme, because it allows the user to fix the within trial variance at specific values per trial.\textsuperscript{23, 24}

Using the output of the model, we obtain the following parameters of interest: (1) the mean of logit sensitivity (\( \theta_A \)) and its standard error; (2) the mean of logit specificity (\( \theta_B \)) and its standard error; (3) the correlation between \( \theta_A \) and \( \theta_B \) given by \( \frac{\sigma_{AB}}{\sqrt{\sigma_A} \cdot \sqrt{\sigma_B}} \); (4) the mean sum of logit sensitivity and specificity (\( \theta_A + \theta_B \)) which is equal to the mean log odds ratio, and its standard error.
3 Direct pooling of sensitivity and specificity using bivariate models in meta-analysis of studies of diagnostic accuracy

Appendix II. Annotated SAS syntax

/*
Original dataset of the 44 studies included in the meta-analysis of Scheidler. Structure of the dataset:
1 record for each study. Key variables: study_id=identification number, modality=type of test 1=CT, 2=LAG, 3=MRI,
and 4 variables holding the number of patients in each of the 4 cells of the 2x2 table
*/
data meta;
input study_id author $ year modality tp fp fn tn;
if tp eq 0 or fp eq 0 or fn eq 0 or tn eq 0 then do;
   tp = tp+0.5;
   fp = fp+0.5;
   fn = fn+0.5;
   tn = tn+0.5;
end;
sens = tp/(tp+fn); spec = tn/(tn+fp); /* calculation of sensitivity and specificity; log_sens = log(sens/(1-sens));
var_log_sens = 1/(sens*(1-sens)*(tp+fn)); * logit transformation with corresponding variance; log_spec = log(spec/(1-spec)); var_log_spec = 1/(spec*(1-spec)*(tn+fp)); */
datalines;
1 Grumbine 1981 1 0 1 6 17
2 Walsh 1981 1 12 3 3 7
3 Brenner 1982 1 4 1 2 13
4 Villasanta 1983 1 10 4 3 25
5 vanEngelshoven 1984 1 3 1 4 12
6 Bandy 1985 1 9 3 3 29
7 Vas 1985 1 20 4 8 31
8 King 1986 1 17 5 7 21
9 Feigen 1987 1 1 1 2 18
10 Camilien 1988 1 3 1 9 38
11 Janus 1989 1 1 1 2 18
12 Matsukuma 1989 1 5 2 2 61
13 Heller 1990 1 21 8 40 184
14 Kim 1990 1 4 3 9 42
15 Ho 1992 1 0 0 5 15
16 Kim 1993 1 7 11 22 158
17 Subak 1995 1 3 3 2 29
18 Kindermann 1970 2 19 1 10 81
19 Lecart 1971 2 8 9 2 13
20 Piver 1971 2 41 1 12 49
21 Piver 1973 2 5 1 2 18
22 Kolbenstvedt 1975 2 45 58 32 165
23 Leman Jr 1975 2 8 6 2 32
24 Brown 1979 2 5 8 1 7
25 Lagasse 1979 2 15 17 11 52
26 Kjorstad 1980 2 16 11 8 24
27 Ashraf 1982 2 4 8 2 25
<table>
<thead>
<tr>
<th>Number</th>
<th>Author</th>
<th>Year</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Disagree</th>
<th>Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>deMuylder</td>
<td>1984</td>
<td>2</td>
<td>8</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>29</td>
<td>Smales</td>
<td>1986</td>
<td>2</td>
<td>10</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>30</td>
<td>Feigen</td>
<td>1987</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>31</td>
<td>Swart</td>
<td>1989</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>32</td>
<td>Heller</td>
<td>1990</td>
<td>2</td>
<td>44</td>
<td>50</td>
<td>12</td>
</tr>
<tr>
<td>33</td>
<td>Laflanze</td>
<td>1990</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>34</td>
<td>Stellato</td>
<td>1992</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>35</td>
<td>Hricak</td>
<td>1988</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>36</td>
<td>Greco</td>
<td>1989</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>37</td>
<td>Janus</td>
<td>1989</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td>Kim</td>
<td>1990</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>39</td>
<td>Ho</td>
<td>1992</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>40</td>
<td>Kim</td>
<td>1993</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>41</td>
<td>Hawnaur</td>
<td>1994</td>
<td>3</td>
<td>12</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>42</td>
<td>Kim</td>
<td>1994</td>
<td>3</td>
<td>23</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>43</td>
<td>Subak</td>
<td>1995</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>44</td>
<td>Heuck</td>
<td>1997</td>
<td>3</td>
<td>16</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

In the next step the original data set is rebuild so that each study now has two records: one record holding (logit) sensitivity and the other (logit) specificity. Two indicator variables, dis and non_dis are made. For the record holding logit sensitivity the value for dis=1 and non_dis=0, for the record holding specificity the values for dis=0 and non_dis=1.

```
data bi_meta;
set meta;
  dis = 1;   non_dis = 0; * creating the record for logit sensitivity ;
  logit = log_sens;   var_logit = var_log_sens;
  rec+1;  output; * variable rec uniquely identifies each record ;
  dis = 0;  non_dis = 1; * creating the record for logit specificity ;
  logit = log_spec;   var_logit = var_log_spec;
  rec+1;  output;
run;
```

Creation of a special dataset holding the 88 variances of logit sensitivity and specificity of each study. Three additional records are created which contain the starting values for the 3 additional variances of the bivariate model: record 1 holds the value for the variance in logit sensitivity between studies, record 2 the variance in logit specificity, record 3 the covariance between logit sensitivity and specificity.
3 Direct pooling of sensitivity and specificity using bivariate models in meta-analysis of studies of diagnostic accuracy

data cov;
if _n_ eq 1 then do;
est = 0.7; output; est = -0.4; output; est = 0.2; output; * the 3 starting values;
end;
set bi_meta; * followed by 88 observed variances of;
est = var_logit; output; * sensitivity and specificity of each study;
keep est; * SAS expects the name est for this variable;
run;
/* Using the proc mixed module in SAS to set up the bivariate model. The approach of Van Houwelingen is used to correctly incorporate both the within and between study variance, more details and explanations can be found there.23 */
proc mixed data=bi_meta method=reml cl ;
class study_id modality; * study_id and modality are categorical variables;
model logit = dis*modality non_dis*modality / noint s cl covb df=1000, 1000, 1000, 1000, 1000, 1000, 1000; * model estimates mean values for sensitivity
and ;
random dis non_dis / subject=study_id type=un; * random effects for logit sensitivity and
specificity;
repeated / group=rec; * using the repeat statement to define the;
 parms / parmdatar=cov hold=4 to 91; * holding the within study variances constant;
contrast 'CT_sens vs LAG_sens' dis*modality 1 -1 0 / df=1000; * testing for differences in sensitivities;
contrast 'CT_sens vs MRI_sens' dis*modality 1 0 -1 / df=1000; * testing for differences in sensitivities;
contrast 'LAG_sens vs MRI_sens' dis*modality 0 1 -1 / df=1000; * testing for differences in sensitivities;
contrast 'CT_spec vs LAG_spec' non_dis*modality 1 -1 0 / df=1000; * testing for differences in specificities;
contrast 'CT_spec vs MRI_spec' non_dis*modality 1 0 -1 / df=1000; * testing for differences in specificities;
contrast 'LAG_spec vs MRI_spec' non_dis*modality 0 1 -1 / df=1000; * testing for differences in specificities;
contrast 'CT_odds vs LAG_odds ' dis*modality 1 -1 0 non_dis*modality 1 -1 0 / df=1000; * testing for differences in
DOR;
contrast 'CT_odds vs MRI_odds ' dis*modality 1 0 -1 non_dis*modality 1 0 -1 / df=1000; * testing for differences in
DOR;
contrast 'LAG_odds vs MRI_odds ' dis*modality 0 1 -1 non_dis*modality 0 1 -1 / df=1000; * testing for differences in
DOR;
run;
4 CT colonography at Ultra-Low Radiation Dose

Abstract

Background - Computed tomography colonography (CTC; virtual colonoscopy) is considered to be a promising method for colorectal cancer screening, but exposure to ionising radiation remains a major drawback. Aim of the study is to evaluate the diagnostic accuracy of CTC at different levels of radiation dose, so to determine the lowest achievable dose without compromising diagnostic accuracy.

Methods - The CTC’s of fifteen patients were simulated to achieve CTC scans equivalent to doses ranging from medium to ultra-low. The sensitivity and specificity were estimated using the free response receiver operating characteristic approach. Conventional colonoscopy was used as the reference standard. In addition inter-observer (2 observers) variability was estimated using the kappa statistic.

Results - The inter-observer variability and detection of clinically relevant polyps were unimpaired down to 2-6% of the medium dose.

Conclusion - Decrease in radiation dose does not impair inter-observer variability and detection of clinical relevant polyps.
Introduction
Computed tomography colonography (CTC; virtual colonoscopy) is becoming an accepted examination for the detection of colorectal polyps and cancers. Since CTC is minimally invasive and highly accurate, it may overcome the patient's reluctance to undergo common colorectal cancer (CRC) screening tests: fecal occult blood test, double contrast barium enema, sigmoidoscopy and colonoscopy. Unfortunately, exposing asymptomatic subjects to radiation doses ranging from 4 to 12 mSv as used today, with an estimated risk of 1 fatal cancer in 10,000 to 3,300 individuals aged 50, will hamper implementation.

As awareness for the potential danger of ionising radiation increases, scanner manufacturers recently introduced CT scanners that enable scans at lower dose than hitherto possible. This development may benefit the use of CTC in CRC screening, as studies have shown that diagnostic accuracy was not impaired at the lowest dose used until now (4 mSv).

The aim of this study was to compare inter-observer variability and polyp detection of CTC at four low to ultra-low dose levels with medium dose CTC in order to determine the lowest dose achievable without compromising accuracy.

Methods
We used CT scans of 15 consecutive surveillance patients (personal or family history of colon polyps or CRC) with at least one clinically relevant polyp (≥5 mm) at colonoscopy. These subjects originated from a comparative study on CTC and colonoscopy after bowel preparation. For the comparative study, the air-distended colon of each patient was scanned with an Mx8000 multi-slice CT scanner (scan parameters: 120 kV, 100 mAs, collimation 4x2.5 mm, rotation time 0.75 s, pitch 1.25, reconstruction interval 1.6 mm) in prone and supine position. The effective dose of a complete examination was 12 mSv. The medical ethics committee approved this study and patients gave written consent.

To compare inter-observer variability and accuracy of CTC at five dose levels in the same fifteen patients, we used an established and validated simulation technique to obtain four additional lower dose CTC exams of each patient. The lowest two doses are not yet achievable in clinical practice, but are technically possible and might be of clinical value.

The simulation method consists of the controlled increase of noise in the original raw CT data. The resulting CT images simulate scans with reduced tube charge (mAs), a parameter that relates proportionally to effective dose. We simulated scans at 25, 6.3, 1.6 and 0.4 mAs, corresponding to effective doses of 3, 0.8, 0.2 and 0.05 mSv, respectively. For the lowest three dose levels the CT images were...
smoothed to obtain approximately constant noise levels, at the expense of spatial resolution.

A radiologist and a senior house officer (experience>100 CTC) unaware of colonoscopic findings and disease prevalence, scored polyps in CTC at the five dose levels in five separate readings. The readings started with the lowest-dose data and subsequently increasing dose levels, at four-week intervals. To further reduce memory bias, the colons were divided in six segments. All segments were evaluated in pairs (prone and supine) in a random order. The majority of segments did not contain polyps. Data were primarily reviewed three-dimensionally with two-dimensional evaluation of suspected lesions.

A polyp detected at CTC was considered to be true positive if it corresponded exactly with the videotaped colonoscopy findings. Inter-observer agreement in terms of segments positive for lesions was determined (k-statistic) for all dose levels. The per patient sensitivity and specificity were estimated using the free response receiver operating characteristic approach.6 7

Results
Ten males and five females were studied, with a mean age of 61 (standard deviation, 13). At colonoscopy 116 polyps were detected; 89 were small (1-5mm), 22 were medium-sized (5-10mm) and 5 were large (≥10mm). K-statistics of CTC for decreasing mAs levels were 0.73, 0.63, 0.68, 0.69 and 0.49, respectively. True and false positive findings of CTC at all mAs levels are listed in table 1.

Detection of polyps of all sizes decreased with decreasing mAs levels but detection of medium to large polyps remained constant for each observer (78-85% and 70-81%), except at 0.4 mAs. Detection of large polyps was comparable at all mAs levels (80-100%). The number of false positive findings decreased with decreasing mAs or remained constant, but at the lowest dose level one observer scored more false positive findings.

Table 2 shows the per patient sensitivity and specificity for both observers. For the detection of smaller polyps these values stay about constant for all but the 0.4 mAs level. For polyps of 10mm or more the diagnostic accuracy is constant.
CT colonography at Ultra-Low Radiation Dose

Table 1 True and false positive results of CT colonography at five mAs levels (100 to 0.4)

<table>
<thead>
<tr>
<th>Lesions according to size</th>
<th>True positive findings in CT colonography</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CS</td>
</tr>
<tr>
<td>All sizes</td>
<td>116</td>
</tr>
<tr>
<td>≥ 5 mm</td>
<td>27</td>
</tr>
<tr>
<td>≥ 10 mm</td>
<td>5</td>
</tr>
<tr>
<td>False positive findings in CT colonography</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td>All sizes</td>
<td>99; 58</td>
</tr>
<tr>
<td>≥ 5 mm</td>
<td>31; 8</td>
</tr>
<tr>
<td>≥ 10 mm</td>
<td>3; 2</td>
</tr>
</tbody>
</table>

The mAs levels are displayed in italic, the results are displayed as radiologist; senior house officer. CS, colonoscopy; CT, computed tomography.

Table 2 Per patient sensitivity (%) and specificity (%) for the two observers (including 95% confidence intervals).

<table>
<thead>
<tr>
<th>mAs</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiologist</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>100 (-)</td>
<td>14 (4.3-36)</td>
<td>100 (-)</td>
<td>33 (15-58)</td>
<td>100 (-)</td>
<td>82 (54-94)</td>
</tr>
<tr>
<td>25</td>
<td>100 (-)</td>
<td>17 (6.3-38)</td>
<td>93 (80-98)</td>
<td>47 (26-69)</td>
<td>80 (31-97)</td>
<td>82 (54-94)</td>
</tr>
<tr>
<td>0.39</td>
<td>96 (87-99)</td>
<td>23 (11-42)</td>
<td>94 (83-99)</td>
<td>42 (22-60)</td>
<td>100 (-)</td>
<td>82 (54-94)</td>
</tr>
</tbody>
</table>

| House officer |              |              |              |              |              |              |
| 25            | 97 (93-100)  | 37 (20-58)  | 91 (76-97)  | 60 (38-79) | 80 (28-98)  | 88 (59-97)  |
| 0.39         | 82 (46-96)  | 43 (1.2-14) | 73 (52-88)  | 74 (47-90) | 80 (28-98)  | 94 (62-99)  |

Discussion
These results indicate that inter-observer variability and detection of clinically relevant polyps remained unimpaired for CTC down to 1.6mAs. Moreover, the number of false positive findings decreased with decreasing mAs. Increased noise and image smoothing in the three lowest dose levels resulted in less precise CT images that may have restrained observers to consider small artefacts to be polyps.

The lowest mAs levels that were tested in this study (1.6 and 0.4), cannot yet be realised in clinical practice. CT scanners that were introduced last year can perform scans at a minimum of 3-5mAs. However, a further reduction of mAs is technically possible and, as this study indicates, clinically useful.

The risk to induce cancer -after a long latent period- due to exposure to
ionising radiation is uncertain. It is, however, generally agreed to aim for the lowest achievable dose. The estimated risk to induce a fatal cancer in a population aged 50 through CTC at 6.3mAs (0.8mSv) is 1 in 55,000, and if 1.6mAs (0.2mSv) becomes available this will be 1 in 220,000.3

This study shows that a dramatic decrease of effective dose for CTC does not impair inter-observer variability and detection of clinically relevant polyps. These findings virtually eliminate the drawback of radiation exposure resulting from CTC and may clear the way for widespread use of this examination in CRC prevention.

References
5 Tumor markers in the diagnosis of primary bladder cancer. A systematic review

Afina S. Glas, Daphne Roos, Marije Deutekom, Aeilko H. Zwinderman, Patrick M.M. Bossuyt, Karl H. Kurth


Abstract

Purpose - To systematically review the available evidence and to obtain and compare summary estimates of the sensitivity and specificity of cytology and the urine based markers bladder tumor antigen, BTA stat, BTA TRAK, NMP 22, telomerase and fibrin degradation products in detecting primary bladder cancer.

Materials and Methods - Studies on the diagnosis of primary bladder cancer published from 1990 through November 2001 in English and German were retrieved from MEDLINE and EMBASE data bases. Included were studies that evaluated one or more of the markers, used cystoscopy as the reference standard and allowed the construction of a 2x2 contingency table for a per patient analysis. The data plus items on study and clinical characteristics were scored by two reviewers. Sensitivity and specificity for each marker were estimated using a bivariate random effect meta analysis. A multivariable analysis was performed to explain between study variation.

Results - A total of 42 studies were included. Only two studies were available on fibrin degradation products, hence a meta-analysis was not possible. Cytology had the best specificity; 94% (95% CI: 90 to 96%). This figure was significantly better than the other markers except for telomerase (specificity 86% (71 to 94%)). Telomerase had the best sensitivity (75% (71 to 79%)) but not significantly better than BTA stat (70% (66-74%)). Case control designs yielded lower values for sensitivity for the tumor markers cytology, bladder tumor antigen and BTA stat.

Conclusions - Cytology has the best specificity and telomerase the best sensitivity. Yet none of the markers studied here is sensitive enough to be recommended for daily routine.
Introduction

Bladder cancer is one of the most common urological cancers. Worldwide it is
diagnosed in about 200,000 men and 60,000 women each year (data for 1990).\textsuperscript{1} The current diagnostic policy starts with screening urine by cytology followed by
cystoscopy and histo-pathological examination of biopsies from the bladder wall.

Cytology from voided urine has been used for detecting bladder cancer in
high risk patients since 1945. It has advantages compared to the expensive and
invasive cystoscopy procedure but lacks the diagnostic sensitivity necessary to rule
out cancer. Since then there have been several attempts to develop easy to use
urinary markers with better sensitivity. This resulted in many urine based marker
tests to detect bladder cancer. Some disappeared after at first promising results.
Others survived and some have entered the market as commercial assay kits.

Although many (comparative) studies have been reported the hierarchy in
diagnostic accuracy of the available markers is not yet clear. The aim of this study
is to obtain and compare summary estimates of the sensitivity and specificity of
the most used urinary tumor markers in detecting primary bladder cancer. Cytology
and the tumor markers bladder tumor antigen (BTA), BTA Stat, BTA TRAK,
NMP 22, telomerase and fibrin degradation products (FDP) were evaluated.
Cytology is the classic marker to which other markers are often compared. BTA,
BTA stat, BTA TRAK, NMP22 and FDP are commercial assays. In addition we
selected telomerase because it might be available soon for physicians. Properties of
the selected markers have been extensively described.\textsuperscript{2} Estimated measures have
been corrected for differences in study characteristics and clinical characteristics if
appropriate.

Methods

Data sources and study selection

The MEDLINE and EMBASE data bases were searched for German and English
articles considering bladder cancer and tumor markers published between January
1990 and November 2001. MeSH headings were used including Bladder
Neoplasm/di,ur, carcinoma/or carcinoma, transitional cell, hematuria, sensitivity
and specificity, rna directed dna polymerase/ or telomerase, fibrogen degradation
products/or fibrogen and urine. An additional search was done using the text
words: dipstick, BTA, NMP22, cytology, marker and urine. Recent reviews were
checked for additional references.

From the abstracts of the identified articles one of the authors (ASG) selected
potentially eligible studies based on the criteria that they incorporated at least one
of the target markers, detection in urine and the indication that sensitivity and specificity were the outcome measures. In case of doubt the article was still retrieved.

All articles on potentially eligible studies were retrieved from the library and independently reviewed by two observers (ASG and DR or MD). Articles were only included if they satisfied the criteria of using cystoscopy and/or histopathology as the reference standard, drawing a 2x2 contingency table, a per patient analysis (per specimen analyses were excluded), diagnosis of primary bladder cancer. If an occasional transitional cell carcinoma of the upper urinary tract or kidney or other type of bladder cancer was detected within a study, the article was not excluded. Disagreements between observers were resolved in a consensus meeting.

Methodological assessment and data extraction
The same readers subsequently extracted data using a standardized score form with elements on methodological quality, clinical characteristics relevant for a particular tumor marker and the outcomes. The criteria for methodological quality were scored and included study design (cohort or case-control), type of control group, a clear description of the study population, reference test and marker test, consecutive patient selection, verification by the reference standard, and independent assessment of the marker test and reference test (blinding). These factors are associated with study quality and have been shown to affect diagnostic test accuracy. Evidence exists that a case-control design can yield overestimated values of sensitivity and specificity relative to a cohort design since often in case control studies a relative sick population is compared with a relatively healthy population. The type of patients selected for the controls (other urological disorders, other patients and/or healthy controls) can effect the specificity. It is expected that healthy controls yield less false positive findings than patients with urological disorders. Verification bias looms when the decision to perform the reference test is based on the result of cytology or urine marker evaluated. This verification is partial when only a sample of the test negatives are subjected to the reference test. Alternatively, this verification is differential if patients with a negative result on the marker test or cytology do not undergo the cystoscopy but are being followed up or get a less invasive reference test. For a case control design these definitions do not apply since diseased and non-diseased populations are selected before the marker test is performed. Interpreting the reference test while knowing the result of cytology or the marker test (nonblind interpretation) can result in an overestimation of the diagnostic accuracy of the test under evaluation. If it was not clear from the article if a blind interpretation of all tests was performed the item was scored as negative. The description of the study population was considered to be sufficient if the age...
distribution, female-to-male ratio and description of the patients signs and symptoms or a description of the stage of patients with bladder cancer was given. The description of the reference and marker test was considered to be sufficient if the cutoff point for positivity and the technique used were given. A consecutive patient sample is needed to prevent selection bias; if it was not evident from the study this item was scored as nonconsecutive. In addition information on bacillus Calmette-Guerin therapy, hematuria, distribution of tumor differentiation of the diseases (grades, WHO classification) and the method of urine collection (voided (including catheterized) or washed), that can influence the outcome of a marker test were scored. Industry sponsorship of a particular study was scored as well since the diagnostic accuracy of a test may be affected.6 Subsequently the raw numbers of the cells of the 2x2 table were retrieved from each article. If not directly presented they were calculated from the sensitivity, specificity or predictive values and the results of the reference test as given. If 2x2 tables of different cut-off values were available, the most commonly used cut-off value was taken. (Matritech 10 U/ml, BTA TRAK 17 U/ml)

Meta-analysis
The results of all studies were plotted in ROC space (receiver operating characteristic curves) to observe the variation between studies. In a ROC space plot the sensitivity of each study is plotted against its 1-specificity. It allows for indication of the between study variation (heterogeneity) and whether this variation is based on threshold differences between studies.7 A study using a threshold of 7 U/ml for the NMP22 can obtain a higher sensitivity and a lower specificity compared to a study using a threshold of 10 U/ml. If the same threshold was used for both studies, the same sensitivity and specificity have been estimated. This effect can also occur for tests with less clear thresholds. A threshold effect can be suspected from the ROC space graph if the points align in a typical shoulder-like pattern. A high correlation between sensitivity and specificity also indicates such a threshold effect.

Analysis was performed using a bivariate random effects approach to pool the sensitivity and specificity for cytology and the tumor markers.8 This model assumes a bivariate normal distribution for the logit-transformed sensitivity and specificity values across studies, allowing for additional heterogeneity between studies due to differences in study characteristics. With this model estimates of the mean logit-transformed sensitivity and specificity values were obtained. Summary estimates of sensitivity and specificity with 95% confidence intervals (CI) were calculated after anti-logarithm transformation of these logit estimates.

To deal with zeroes in the 2x2 table 0.5 was added to all counts.7
A multivariable analysis was done to explain variation in sensitivity and specificity between studies. Covariates were selected if a specific methodological or clinical variable correlated (Spearman) with sensitivity or specificity with a p value smaller than 0.1. Statistical analyses were executed with the statistical software packages: SPSS version 10, SPSS Inc. Chicago, IL, USA and SAS version 8.02, SAS institute Inc. Cary, NC, USA.

Results

Search results, methodological assessment and data extraction
Based on the inclusion criteria 133 potentially eligible articles were identified, of which 91 could not be included with 15 drawing 2x2 table not possible, 7 with only data on sensitivity or specificity, 8 with data per specimen not per patient, 13 with data on primary and follow-up with no possibility to split up the data, 21 only follow up, 13 reviews, 1 suspected duplicate publication and 13 without the target urine marker or only dipstick.

Forty-two articles were suitable for the analysis. Table 1 lists these articles together with information about the year of publication, the urinebased markers studied, design of the study, the grades (WHO classification) of the diseased and the condition of the non-diseased (urological patients, other patients or both). In addition 8 (19%, 8/42) studies mentioned explicitly that they had been sponsored by the industry. All but one study with a cohort design verified all their patients by the reference standard. In the one study only patients with a positive marker test were subjected to the reference standard, with test negatives verified by follow up. A clear description of the demographics of the patients, the reference test and the marker test was found in respectively 40%, 7% and 86% of the studies. A blind interpretation of the studies and the use of a consecutive sample was only alleged from 33% and 11% of the studies respectively.
Table 1: Study characteristics of the 42 included studies.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Marker*</th>
<th>Design</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Controls ‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planz et al.</td>
<td>2000</td>
<td>1</td>
<td>case control</td>
<td>68</td>
<td>25</td>
<td>7.5</td>
<td>both</td>
</tr>
<tr>
<td>Ito et al.</td>
<td>1998</td>
<td>1,6</td>
<td>case control</td>
<td>27</td>
<td>55</td>
<td>18</td>
<td>100% uro</td>
</tr>
<tr>
<td>Rahat et al.</td>
<td>1999</td>
<td>6</td>
<td>case control</td>
<td>12</td>
<td>55</td>
<td>18</td>
<td>both</td>
</tr>
<tr>
<td>Kaieter et al.</td>
<td>1998</td>
<td>6</td>
<td>case control</td>
<td>28</td>
<td>37</td>
<td>31</td>
<td>100% uro</td>
</tr>
<tr>
<td>Yoshida et al.</td>
<td>1997</td>
<td>6</td>
<td>case control</td>
<td>27</td>
<td>38</td>
<td>31</td>
<td>70% uro</td>
</tr>
<tr>
<td>Kirollos et al.</td>
<td>1997</td>
<td>1,2</td>
<td>cohort</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sozen et al.</td>
<td>1999</td>
<td>1,3,5</td>
<td>case control</td>
<td>35</td>
<td>40</td>
<td>25</td>
<td>100% uro</td>
</tr>
<tr>
<td>Heicapell et al.</td>
<td>1999</td>
<td>4</td>
<td>case control</td>
<td>7.7</td>
<td>43</td>
<td>48</td>
<td>60% uro</td>
</tr>
<tr>
<td>Leh et al.</td>
<td>1999</td>
<td>1,3</td>
<td>cohort</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zippe et al.</td>
<td>1999</td>
<td>1,5</td>
<td>cohort</td>
<td>11</td>
<td>33</td>
<td>56</td>
<td>NA</td>
</tr>
<tr>
<td>Mian et al.</td>
<td>1999</td>
<td>1</td>
<td>cohort</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Johnston et al.</td>
<td>1997</td>
<td>1,2,7</td>
<td>case control</td>
<td>37</td>
<td>39</td>
<td>33</td>
<td>100% uro</td>
</tr>
<tr>
<td>Murphy et al.</td>
<td>1997</td>
<td>1,2</td>
<td>case control</td>
<td>14</td>
<td>21</td>
<td>39</td>
<td>100% uro</td>
</tr>
<tr>
<td>Casella et al.</td>
<td>2000</td>
<td>1,5</td>
<td>both</td>
<td>32</td>
<td>38</td>
<td>29</td>
<td>100% uro</td>
</tr>
<tr>
<td>Sharma et al.</td>
<td>1999</td>
<td>1,3,5</td>
<td>cohort</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ramakuram et al.</td>
<td>1999</td>
<td>1,3,5,6,7</td>
<td>case control</td>
<td>16</td>
<td>46</td>
<td>23</td>
<td>0% uro</td>
</tr>
<tr>
<td>Landman et al.</td>
<td>1998</td>
<td>12,5,6</td>
<td>case control</td>
<td>34</td>
<td>31</td>
<td>33</td>
<td>100% uro</td>
</tr>
<tr>
<td>Kinoshita et al.</td>
<td>1997</td>
<td>6</td>
<td>case control</td>
<td>27</td>
<td>62</td>
<td>11</td>
<td>100% uro</td>
</tr>
<tr>
<td>Thomas et al.</td>
<td>1999</td>
<td>4</td>
<td>cohort</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gelmini et al.</td>
<td>2000</td>
<td>6</td>
<td>case control</td>
<td>33</td>
<td>15</td>
<td>52</td>
<td>0% uro</td>
</tr>
<tr>
<td>Leh et al.</td>
<td>1997</td>
<td>2</td>
<td>cohort</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chong et al.</td>
<td>1999</td>
<td>2</td>
<td>cohort</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heicapell et al.</td>
<td>2000</td>
<td>3</td>
<td>case control</td>
<td>13</td>
<td>40</td>
<td>45</td>
<td>60% uro</td>
</tr>
<tr>
<td>Giannopoulos et al.</td>
<td>2001</td>
<td>3,5</td>
<td>cohort</td>
<td>19</td>
<td>40</td>
<td>41</td>
<td>NA</td>
</tr>
<tr>
<td>Pode et al.</td>
<td>1999</td>
<td>1,3</td>
<td>cohort</td>
<td>7</td>
<td>44</td>
<td>49</td>
<td>NA</td>
</tr>
<tr>
<td>Miyanga et al.</td>
<td>1999</td>
<td>1,5</td>
<td>cohort</td>
<td>14</td>
<td>36</td>
<td>50</td>
<td>NA</td>
</tr>
<tr>
<td>Nasuti et al.</td>
<td>1999</td>
<td>1,3</td>
<td>cohort</td>
<td>0</td>
<td>66</td>
<td>33</td>
<td>NA</td>
</tr>
<tr>
<td>Paoluzzi et al.</td>
<td>1999</td>
<td>1,5</td>
<td>cohort</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abbate et al.</td>
<td>1998</td>
<td>4,5</td>
<td>case control</td>
<td>25</td>
<td>29</td>
<td>46</td>
<td>36% uro</td>
</tr>
<tr>
<td>Mahner et al.</td>
<td>1999</td>
<td>4</td>
<td>case control</td>
<td>NA</td>
<td></td>
<td></td>
<td>0% uro</td>
</tr>
<tr>
<td>Cajulis et al.</td>
<td>1994</td>
<td>1</td>
<td>cohort</td>
<td>12</td>
<td>23</td>
<td>65</td>
<td>NA</td>
</tr>
<tr>
<td>Casetta et al.</td>
<td>2000</td>
<td>1</td>
<td>cohort</td>
<td>26</td>
<td>38</td>
<td>36</td>
<td>NA</td>
</tr>
<tr>
<td>Lee et al.</td>
<td>2001</td>
<td>1,5</td>
<td>case control</td>
<td>27</td>
<td>49</td>
<td>24</td>
<td>0% uro</td>
</tr>
<tr>
<td>Mondal et al.</td>
<td>1992</td>
<td>1</td>
<td>cohort</td>
<td>17</td>
<td>14</td>
<td>57</td>
<td>NA</td>
</tr>
<tr>
<td>Cheng et al.</td>
<td>2000</td>
<td>6</td>
<td>case control</td>
<td>23</td>
<td>41</td>
<td>29</td>
<td>100% uro</td>
</tr>
<tr>
<td>O'Donoghue et al.</td>
<td>1991</td>
<td>1</td>
<td>cohort</td>
<td>27</td>
<td>25</td>
<td>16</td>
<td>NA</td>
</tr>
<tr>
<td>Zhang et al.</td>
<td>2001</td>
<td>1</td>
<td>case control</td>
<td>29</td>
<td>47</td>
<td>24</td>
<td>60% uro</td>
</tr>
<tr>
<td>Chahal et al.</td>
<td>2001</td>
<td>1,5</td>
<td>cohort</td>
<td>38</td>
<td>19</td>
<td>44</td>
<td>NA</td>
</tr>
<tr>
<td>Priola et al.</td>
<td>2001</td>
<td>1,4</td>
<td>cohort</td>
<td>29</td>
<td>37</td>
<td>34</td>
<td>NA</td>
</tr>
<tr>
<td>Oge et al.</td>
<td>2001</td>
<td>5</td>
<td>cohort</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cassel et al.</td>
<td>2001</td>
<td>6</td>
<td>case control</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lahme et al.</td>
<td>2001</td>
<td>1,5</td>
<td>case control</td>
<td>NA</td>
<td></td>
<td></td>
<td>19% uro</td>
</tr>
</tbody>
</table>

*) 1: cytology, 2: BTA, 3: BTA Stat, 4: BTA TRAK, 5: NMP22, 6: Telomerase, 7: FDP. † Grades: If percentages do not add up to 100% the residual percentage is due to CIS or other tumor histology. ‡ Controls: The types of controls selected in case of a case control design: NA = not applicable, % uro = percentage patients with urological conditions other than bladder cancer. If not 100% the other controls are non-urological patients and/or healthy controls. Both = from the particular study no percentages could be calculated.
Table 2 and Table 3 show the studies evaluating cytology and each urinary marker with the raw numbers as extracted from the individual studies. The sensitivity and specificity given are calculated from the raw numbers. Table 3 also contains the cut-off values used for quantitative markers.

Table 2 Individual studies evaluating cytology.

<table>
<thead>
<tr>
<th>Author (first)</th>
<th>Urine</th>
<th>No</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Sensitivity %</th>
<th>Specificity %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planz</td>
<td>Voided</td>
<td>80</td>
<td>19</td>
<td>0</td>
<td>21</td>
<td>40</td>
<td>48</td>
<td>100</td>
</tr>
<tr>
<td>Ito</td>
<td>Voided</td>
<td>59</td>
<td>22</td>
<td>0</td>
<td>11</td>
<td>26</td>
<td>67</td>
<td>100</td>
</tr>
<tr>
<td>Kirollos</td>
<td>Voided</td>
<td>54</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>52</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Sozen</td>
<td>Voided</td>
<td>140</td>
<td>14</td>
<td>10</td>
<td>26</td>
<td>90</td>
<td>35</td>
<td>90</td>
</tr>
<tr>
<td>Leyh</td>
<td>Both</td>
<td>105</td>
<td>18</td>
<td>0</td>
<td>37</td>
<td>50</td>
<td>33</td>
<td>100</td>
</tr>
<tr>
<td>Zippe</td>
<td>Voided</td>
<td>330</td>
<td>6</td>
<td>0</td>
<td>12</td>
<td>312</td>
<td>33</td>
<td>100</td>
</tr>
<tr>
<td>Mian</td>
<td>Voided</td>
<td>107</td>
<td>14</td>
<td>1</td>
<td>9</td>
<td>83</td>
<td>61</td>
<td>99</td>
</tr>
<tr>
<td>Johnston</td>
<td>Voided</td>
<td>130</td>
<td>21</td>
<td>7</td>
<td>39</td>
<td>63</td>
<td>35</td>
<td>90</td>
</tr>
<tr>
<td>Murphy</td>
<td>Voided</td>
<td>62</td>
<td>15</td>
<td>0</td>
<td>13</td>
<td>34</td>
<td>54</td>
<td>100</td>
</tr>
<tr>
<td>Casella</td>
<td>Washed</td>
<td>200</td>
<td>50</td>
<td>11</td>
<td>45</td>
<td>94</td>
<td>53</td>
<td>90</td>
</tr>
<tr>
<td>Sharma</td>
<td>Voided</td>
<td>199</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>185</td>
<td>33</td>
<td>96</td>
</tr>
<tr>
<td>Ramakumar</td>
<td>Voided</td>
<td>113</td>
<td>24</td>
<td>4</td>
<td>30</td>
<td>55</td>
<td>44</td>
<td>93</td>
</tr>
<tr>
<td>Landman</td>
<td>Voided</td>
<td>77</td>
<td>19</td>
<td>2</td>
<td>28</td>
<td>28</td>
<td>40</td>
<td>93</td>
</tr>
<tr>
<td>Pode</td>
<td>Voided</td>
<td>85</td>
<td>38</td>
<td>3</td>
<td>33</td>
<td>11</td>
<td>54</td>
<td>79</td>
</tr>
<tr>
<td>Miyanaga</td>
<td>Voided</td>
<td>309</td>
<td>12</td>
<td>1</td>
<td>10</td>
<td>286</td>
<td>55</td>
<td>100</td>
</tr>
<tr>
<td>Nasuti</td>
<td>Voided</td>
<td>100</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>93</td>
<td>100</td>
<td>96</td>
</tr>
<tr>
<td>Paoluzzi</td>
<td>Voided</td>
<td>90</td>
<td>24</td>
<td>22</td>
<td>8</td>
<td>36</td>
<td>75</td>
<td>62</td>
</tr>
<tr>
<td>Cajulis</td>
<td>Washed</td>
<td>40</td>
<td>16</td>
<td>0</td>
<td>10</td>
<td>14</td>
<td>62</td>
<td>100</td>
</tr>
<tr>
<td>Casetta</td>
<td>Voided</td>
<td>94</td>
<td>48</td>
<td>1</td>
<td>32</td>
<td>13</td>
<td>60</td>
<td>93</td>
</tr>
<tr>
<td>Lee</td>
<td>Voided</td>
<td>106</td>
<td>39</td>
<td>4</td>
<td>31</td>
<td>32</td>
<td>56</td>
<td>89</td>
</tr>
<tr>
<td>Mondal</td>
<td>Voided</td>
<td>139</td>
<td>79</td>
<td>4</td>
<td>5</td>
<td>51</td>
<td>94</td>
<td>93</td>
</tr>
<tr>
<td>O'Donoghue</td>
<td>Voided</td>
<td>265</td>
<td>42</td>
<td>5</td>
<td>9</td>
<td>209</td>
<td>82</td>
<td>98</td>
</tr>
<tr>
<td>Zhang</td>
<td>Voided</td>
<td>101</td>
<td>41</td>
<td>7</td>
<td>40</td>
<td>13</td>
<td>51</td>
<td>65</td>
</tr>
<tr>
<td>Chahal</td>
<td>Voided</td>
<td>96</td>
<td>5</td>
<td>4</td>
<td>11</td>
<td>76</td>
<td>31</td>
<td>95</td>
</tr>
<tr>
<td>Priolo</td>
<td>Voided</td>
<td>194</td>
<td>105</td>
<td>12</td>
<td>42</td>
<td>35</td>
<td>71</td>
<td>74</td>
</tr>
<tr>
<td>Lahme</td>
<td>Voided</td>
<td>169</td>
<td>18</td>
<td>5</td>
<td>22</td>
<td>124</td>
<td>45</td>
<td>96</td>
</tr>
</tbody>
</table>

*) Catheterized urine is also scored as voided. † † No: the number of patients per original study. † † TP: true positives, FP: false positives, FN: false negatives, TN: true negatives.
5 Tumor markers in the diagnosis of primary bladder cancer. A systematic review

Table 3 Individual studies evaluating urine markers: raw numbers and outcome

<table>
<thead>
<tr>
<th>Marker</th>
<th>Author</th>
<th>Cutoff U/ml</th>
<th>No</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Sensitivity %</th>
<th>Specificity %</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTA</td>
<td>Kirollos</td>
<td>54</td>
<td>15</td>
<td>1</td>
<td>37</td>
<td>50</td>
<td>71</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Johnston</td>
<td>130</td>
<td>17</td>
<td>9</td>
<td>43</td>
<td>61</td>
<td>28</td>
<td>71</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>Murphy</td>
<td>62</td>
<td>8</td>
<td>7</td>
<td>20</td>
<td>27</td>
<td>29</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Landman</td>
<td>77</td>
<td>19</td>
<td>8</td>
<td>28</td>
<td>22</td>
<td>40</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leyh</td>
<td>345</td>
<td>33</td>
<td>27</td>
<td>8</td>
<td>277</td>
<td>80</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chong</td>
<td>47</td>
<td>8</td>
<td>12</td>
<td>4</td>
<td>23</td>
<td>67</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Johnston</td>
<td>140</td>
<td>28</td>
<td>32</td>
<td>12</td>
<td>68</td>
<td>70</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Johnston</td>
<td>105</td>
<td>40</td>
<td>24</td>
<td>15</td>
<td>26</td>
<td>73</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sharma</td>
<td>199</td>
<td>8</td>
<td>7</td>
<td>20</td>
<td>27</td>
<td>29</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ramakumar</td>
<td>196</td>
<td>42</td>
<td>15</td>
<td>101</td>
<td>74</td>
<td>73</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Heicapell</td>
<td>354</td>
<td>105</td>
<td>13</td>
<td>62</td>
<td>174</td>
<td>63</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Giannopoulos</td>
<td>118</td>
<td>50</td>
<td>22</td>
<td>18</td>
<td>28</td>
<td>74</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pode</td>
<td>88</td>
<td>64</td>
<td>7</td>
<td>13</td>
<td>90</td>
<td>76</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nasuti</td>
<td>100</td>
<td>3</td>
<td>16</td>
<td>0</td>
<td>81</td>
<td>100</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>BTA stat</td>
<td>Heicapell</td>
<td>17</td>
<td>298</td>
<td>75</td>
<td>96</td>
<td>29</td>
<td>98</td>
<td>72</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Thomas</td>
<td>14</td>
<td>9</td>
<td>57</td>
<td>22</td>
<td>12</td>
<td>25</td>
<td>76</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Abbate</td>
<td>23</td>
<td>107</td>
<td>33</td>
<td>1</td>
<td>20</td>
<td>53</td>
<td>62</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Mahnert</td>
<td>14</td>
<td>134</td>
<td>47</td>
<td>27</td>
<td>29</td>
<td>31</td>
<td>62</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Priolo</td>
<td>35</td>
<td>194</td>
<td>94</td>
<td>17</td>
<td>53</td>
<td>30</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Sozen</td>
<td>10</td>
<td>140</td>
<td>29</td>
<td>19</td>
<td>11</td>
<td>81</td>
<td>73</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>BTA TRAK</td>
<td>10</td>
<td>330</td>
<td>18</td>
<td>45</td>
<td>0</td>
<td>267</td>
<td>100</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>Casella</td>
<td>10</td>
<td>235</td>
<td>67</td>
<td>17</td>
<td>63</td>
<td>88</td>
<td>52</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>Sharma</td>
<td>10</td>
<td>205</td>
<td>4</td>
<td>33</td>
<td>2</td>
<td>166</td>
<td>67</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>Ramakumar</td>
<td>10</td>
<td>196</td>
<td>56</td>
<td>27</td>
<td>83</td>
<td>53</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Landman</td>
<td>7</td>
<td>77</td>
<td>38</td>
<td>7</td>
<td>9</td>
<td>23</td>
<td>81</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Giannopoulos</td>
<td>8</td>
<td>118</td>
<td>47</td>
<td>16</td>
<td>21</td>
<td>34</td>
<td>69</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>Miyakage</td>
<td>12</td>
<td>309</td>
<td>68</td>
<td>2</td>
<td>219</td>
<td>91</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paoluzzi</td>
<td>10</td>
<td>90</td>
<td>27</td>
<td>2</td>
<td>5</td>
<td>36</td>
<td>84</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>Abbate</td>
<td>12</td>
<td>182</td>
<td>59</td>
<td>4</td>
<td>50</td>
<td>69</td>
<td>54</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>Lee</td>
<td>7,7</td>
<td>106</td>
<td>53</td>
<td>10</td>
<td>17</td>
<td>26</td>
<td>76</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>Chahal</td>
<td>10</td>
<td>96</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>73</td>
<td>44</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Lahme</td>
<td>10</td>
<td>169</td>
<td>25</td>
<td>31</td>
<td>15</td>
<td>98</td>
<td>63</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Öge</td>
<td>10</td>
<td>37</td>
<td>20</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>74</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Itto</td>
<td>10</td>
<td>59</td>
<td>25</td>
<td>1</td>
<td>8</td>
<td>25</td>
<td>76</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>Rahat</td>
<td>35</td>
<td>17</td>
<td>3</td>
<td>4</td>
<td>11</td>
<td>81</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kavalier</td>
<td>151</td>
<td>88</td>
<td>16</td>
<td>16</td>
<td>31</td>
<td>85</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yoshida</td>
<td>109</td>
<td>16</td>
<td>3</td>
<td>10</td>
<td>80</td>
<td>62</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ramakumar</td>
<td>195</td>
<td>40</td>
<td>1</td>
<td>17</td>
<td>137</td>
<td>70</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Landman</td>
<td>77</td>
<td>38</td>
<td>6</td>
<td>9</td>
<td>24</td>
<td>81</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kinoshita</td>
<td>54</td>
<td>23</td>
<td>0</td>
<td>19</td>
<td>12</td>
<td>55</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gelmini</td>
<td>53</td>
<td>27</td>
<td>2</td>
<td>6</td>
<td>18</td>
<td>82</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cheng</td>
<td>49</td>
<td>14</td>
<td>3</td>
<td>3</td>
<td>29</td>
<td>82</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cassel</td>
<td>73</td>
<td>37</td>
<td>22</td>
<td>7</td>
<td>7</td>
<td>84</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Johnstone</td>
<td>130</td>
<td>47</td>
<td>17</td>
<td>13</td>
<td>53</td>
<td>78</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ramakumar</td>
<td>186</td>
<td>27</td>
<td>12</td>
<td>25</td>
<td>122</td>
<td>52</td>
<td>91</td>
<td></td>
</tr>
</tbody>
</table>
(Meta) analysis

Figure 1 shows the ROC space plots for each urinary marker. It can be observed that sensitivity is more heterogeneous than specificity for cytology, BTA and NMP22. For BTA TRAK and telomerase it is the opposite, yielding a small range in sensitivity. Only for telomerase a threshold affect seems apparent, meaning that part of the variation between studies is due to differences in threshold. This finding is supported by the significant correlation between the logit sensitivity and logit specificity (table 4).

![Figure 1](image-url)
The results of the overall bivariate analysis are listed in table 4. Cytology had a pooled specificity of 94% (95% CI: 90 to 96%). This is statistically significant higher than the specificity of the BTA, BTA Stat, BTA TRAK and NMP22 test (see also figure 2). Cytology and telomerase were not significantly different as tested with the z-test (p=0.13). Telomerase had the best sensitivity, statistically different from NMP22 (p=0.04) and the other markers, but not statistically significant from BTA stat (p=0.13). Only two studies were eligible for the FDP test, so pooling these two studies was not feasible.

A multivariable analysis was performed per marker to identify factors that could explain additional variability between studies. The influence of hematuria could not be scored from most articles. For cytology sensitivity correlated with year of publication and the design used (both p<0.1). The sensitivity of case control designs (48%, 95% CI: 39 to 57%) was lower than of cohort designs (61%, 95% CI: 52 to 69%). (p=0.04) Between 1990 and 2000 the sensitivity decreased linearly from 80 to 52% and the specificity decreased from 97 to 92%. Specificity did not correlate with any of the scored variables. Study design did also effect the sensitivity of the BTA test. A case control design yielded significantly lower estimates (33%, CI: 26 to 41%) than a cohort study (73%, CI: 60 to 83%). The specificity of the BTA stat test was correlated with the manner of test interpretation, blind or not blind from other information. Studies with blind interpretation had a lower specificity (59% (CI: 46 to 71%)) than studies with a non-blind interpretation of the test results (83% (CI: 76 to 88)). Sensitivity of BTA stat was correlated with design. The sensitivity for the BTA stat marker was 66 (CI: 60 to 71%) for a case control design and 77% (CI: 71 to 82) for a cohort design. The NMP22 marker had a positive correlation of both sensitivity and specificity with the method of patients selection (p<0.1), but this correlation was based on only two studies with a sensitivity of respectively 54 and 44% and a specificity of respectively 95 and 91%. After exploring the position of these two studies in figure 1 (left under), a separate multivariable analysis based on this variable was judged not appropriate. All studies evaluating telomerase were of the case control design, thus no effect of design on the accuracy for research on this particular marker could be studied. The type of control group did not significantly affect the specificity, as was the case for all markers. Only publication year correlated with sensitivity and specificity (p<0.1). From 1997 until 2001 the average sensitivity increased from 67 to 95% and the average specificity decreased from 95 to 62%. For the BTA TRAK neither sensitivity nor specificity correlated with any of the variables.
Discussion
We systematically reviewed the available evidence of eight urine based markers for the detection of primary bladder cancer. Cytology was found to have the best specificity and telomerase the best sensitivity. The results of FDP could not be summarized since only 2 studies suited our inclusion criteria. The multivariable analysis of cytology, BTA and BTA stat revealed that a part of the variation in sensitivity within studies could be explained by the study design used. A case control design yielded lower estimates of sensitivity than a cohort design. For cytology the year of publication influenced the sensitivity, while the specificity remained about constant. On the other hand the sensitivity of telomerase increased in time while the specificity decreased in time. Furthermore a nonblind interpretation of test result gave an overestimation of the specificity for BTA stat.

Table 4 The results of the bivariate analysis.

<table>
<thead>
<tr>
<th>Marker</th>
<th>Sensitivity % (95% CI)</th>
<th>Specificity % (95% CI)</th>
<th>No*</th>
<th>Correlation†</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytology</td>
<td>55 (48 to 62)</td>
<td>94 (90 to 96)</td>
<td>3444</td>
<td>-0.11</td>
<td>0.6</td>
</tr>
<tr>
<td>BTA</td>
<td>50 (30 to 65)</td>
<td>79 (70 to 86)</td>
<td>661</td>
<td>0.12</td>
<td>0.7</td>
</tr>
<tr>
<td>BTA stat</td>
<td>70 (66 to 74)</td>
<td>75 (64 to 84)</td>
<td>1160</td>
<td>-0.092</td>
<td>0.8</td>
</tr>
<tr>
<td>BTA TRAK</td>
<td>66 (62 to 71)</td>
<td>65 (45 to 81)</td>
<td>829</td>
<td>-0.50</td>
<td>0.4</td>
</tr>
<tr>
<td>NMP22</td>
<td>67 (60 to 73)</td>
<td>78 (72 to 83)</td>
<td>2290</td>
<td>-0.15</td>
<td>0.6</td>
</tr>
<tr>
<td>Telomerase</td>
<td>75 (71 to 79)</td>
<td>86 (71 to 94)</td>
<td>855</td>
<td>-0.73</td>
<td>0.02</td>
</tr>
</tbody>
</table>

* total number of patients
† Pearson's correlation coefficient between the logit sensitivity and logit specificity

Figure 2 Summary estimates of sensitivity and specificity for each tumor marker with the 95% CI. Note that the X-axes start at 25, to increase the readability of the graph.
5 Tumor markers in the diagnosis of primary bladder cancer. A systematic review

The results of the overall analyses are comparable with the estimates published in a recent review. In that review the literature was not systematically searched and it was not reported how the estimates were pooled. Since it is not clear on what basis articles were selected and whether study size or threshold differences were taken into account comparisons based on such reviews can be dicey. Available head to head comparisons support our observation that cytology has best specificity and telomerase has best sensitivity. Our finding that a case control design yielded lower values for sensitivity contradicts the opinion that case control studies generally overestimate results. An explanation for the excess false negatives in the case control design relative to a cohort design might be that the patients sampled had relatively low grade cancers. Another explanation might be that the selected patients already received therapy. An attempt was made to retrieve information on given therapy to cases, but the original studies reported poorly on this.

This evaluation deals with urine based markers used for the detection of primary bladder cancers. Yet these tumor markers are mostly used in adjunct to cystoscopy in the surveillance of patients at high risk for bladder cancer, for example patients with a history of this disease. Hence the results of this study should not be extrapolated to cases, but should be evaluated separately.

We used a bivariate random effects approach for summarizing estimates of accuracy. This method is more convenient than the model described by Littenberg and colleagues, which has a summary receiver operating characteristic (SROC) curve as outcome. The outcomes of the bivariate approach are the summary estimates of sensitivity and specificity, which can be interpreted as a pair. These outcomes are more familiar to clinicians. From the SROC curve also sensitivity and specificity can be estimated but this is not clear cut, since along the SROC curve many different pairs of sensitivity and specificity can be chosen. An additional advantage of this bivariate random effects model is that its random effects nature allows both systematical and coincidental differences between studies.

One of the limitations of all meta-analyses is that the quality of the original studies affects the validity of the final results. The quality of the original studies included here was weak. Only a few studies could be identified as having a consecutive series of patients suspected for bladder cancer with independent assessment of the marker test and reference standard. These criteria are the keystones to conduct correct diagnostic test research. Another limitation is that few studies were available for the BTA and BTA related markers as compared to cytology, NMP22 and telomerase. This leads to less precise estimates with broad confidence intervals. These might be the reason that no differences were found in sensitivity between telomerase and BTA stat and telomerase and NMP22 as well as in...
specificity between cytology and telomerase. One more limitation was that the sensitivity and specificity in subgroups of tumor grade and stage could not be evaluated.

What marker is most suitable for use in clinical practice? The aim of developing urine based markers in diagnosing bladder cancer is to triage patients with symptoms suspected for the disease before undergoing ‘invasive’ cystoscopy. Thus a very high sensitivity (near 100%) is needed so that practically no patients with bladder cancer will be missed and preferable a high specificity to prevent that patients unnecessarily undergo cystoscopy. In addition it should be non-invasive, fast and low in costs. It is shown here that on average none of the markers has a sensitivity over 80%. This is not high enough for clinical use, no matter how low the costs and/or how fast the procedure time of these markers.

The sensitivity of telomerase was estimated in case control studies only. In this meta-analysis case control studies were found to yield lower values of sensitivity relative to cohort studies. Thus it is possible that telomerase will have a better sensitivity if evaluated in a cohort of patients suspected for primary bladder cancer, representing the clinical situation. Cytology was found to have the best specificity. A combination of cytology and telomerase might thus be suitable for the triage of patients with symptoms of bladder cancer, but we have no data from the current meta-analysis to evaluate specific test combinations.

References
5 Tumor markers in the diagnosis of primary bladder cancer. A systematic review

5 Tumor markers in the diagnosis of primary bladder cancer. A systematic review


6 The added value of pre-discharge Dobutamine Stress Echocardiography in the long-term risk stratification of chest pain patients

Afina S. Glas, Radha Bholasingh, Petrus M. van der Zee, Jeroen G. Lijmer, Patrick M.M. Bossuyt, Robbert J. de Winter.

Abstract

Background - Dobutamine stress echocardiography (DSE) has been proposed as a risk stratification tool in patients with chest complaints with a normal or non-diagnostic electrocardiogram on presentation at the emergency room.

Objective - To evaluate the added value of DSE in the long-term risk assessment relative to physicians’ risk estimates.

Methods - Chest pain patients with a normal or non-diagnostic ECG presenting at the emergency room were included after ruling out acute coronary syndrome. At admission and at discharge attending physicians estimated patient-specific probabilities of a cardiac event in the next 6 months of which the calibration was evaluated. A pre-discharge DSE was performed, of which the results were not disclosed to the physicians. The DSE likelihood ratios were used to transform physicians’ discharge estimates (pre-test probabilities) into post-test probabilities. The main outcome measure used was the area under the receiver operating characteristic (ROC) curve to compare pretest probabilities with posttest probabilities.

Results - The 6-month cardiac event rate in the 398 included patients was 7.5%. DSE likelihood ratios were 7.8 (95% CI: 4.2 to 14) for a positive and 0.63 (0.47 to 1.8) for a negative result. Physicians’ probabilities were well calibrated. The area under the ROC curve did not significantly improve by including the DSE results.

Conclusion - Due to the low event rate and an undiscriminating negative likelihood ratio the DSE has limited additional value in the long term risk stratification of chest pain patients, with a possible exception for patients judged to be at intermediate risk.
The added value of pre-discharge Dobutamine Stress Echocardiography in the long-term risk stratification of chest pain patients

Introduction
Patients presenting with chest pain at the emergency department constitute a heterogeneous population as to the cause of their symptoms. Diagnosis varies from atypical chest pain to acute myocardial infarction. The short-term risk stratification aims at identification of urgent conditions requiring immediate treatment: acute coronary syndromes. These patients will be admitted and further management will be relatively straightforward.

The majority of chest pain patients presenting at the emergency room, however, do not have an acute coronary syndrome, but nevertheless are at increased risk of subsequent cardiac events. Reported long-term cardiac rates in these patient groups vary from 6 to 12%.\(^1\)\(^2\) These disturbingly high risks require further investigations in these patients, aimed at more accurate risk assessment, identifying patients with an acceptably low risk of a cardiac event, whom can be safely discharged without follow-up.

Several objective diagnostic tools are available for long-term risk assessment. An abnormal exercise test is an indicator of poor prognosis, but the value of this test is limited, especially in patients with an non-interpretable electrocardiogram (LVH, BBB, pacemaker, medication). An exercise test is often impracticable due to concomitant conditions such as peripheral artery disease, obstructive pulmonary disease, and neurological or muscular-skeletal disorders. Vasodilator stress testing with either dipyridamole or adenosine in conjunction with thallium/MIBI perfusion imaging or echocardiogram may have better accuracy, but is relatively laborious and may not be readily available for patients presenting at the emergency room.

An attractive alternative is the dobutamine stress echocardiography (DSE). The diagnostic and prognostic value of DSE has been demonstrated in patients with known or suspected coronary artery disease, post myocardial infarction, and prior to non-cardiac vascular surgery. Pre-discharge dobutamine stress echocardiography has been shown to carry independent prognostic value in low risk, troponin negative chest pain patients.\(^3\)

Despite this well-demonstrated association between the results of pre-discharge dobutamine stress echocardiography and cardiac events, one can question whether such a test is indicated in all patients. Patients presenting at the emergency room do not form a homogeneous population, due to variability in complaints, history and presentation, which leads to differences in the degree of suspicion with respect to future events.

For this purpose we evaluated the added value of DSE in the long-term risk stratification of patients with chest complaints with an inconclusive electrocardiogram and ruled out acute coronary syndrome. The added value is defined relative to the
degree of suspicion expressed by the physician, based on information available, expressed in a subjective probability. Our aim was to identify patients at low risk for a cardiac event after discharge from the emergency room, after they had been admitted with typical chest pain. The reference standard was the blindly assessed occurrence of a cardiac event within 6 months after admission.

Patients and Methods

Patients
Data were obtained as part of a prospective study that was performed from July 1997 to April 2000 at the cardiac emergency room of three Dutch hospitals.3 Consecutive patients with typical chest pain within 6 hours of onset, with a normal or non-diagnostic electrocardiogram where included if they were at least 18 years and had given written informed consent. Acute myocardial infarction (AMI) and unstable angina pectoris were ruled out in a standard protocol of 12 hour observation. AMI was defined by a rise of CKMB levels of twice the upper limit of normal or more. Unstable angina was defined as repetitive episodes of chest pain accompanied by ECG changes despite optimal oral medication and/or requiring i.v. medication. Excluded were patients with ventricular tachycardia, chronic atrial fibrillation, conduction disturbances, severe uncontrolled hypertension (systolic >180 mmHg or diastolic > 120 mmHg despite therapy), overt heart failure requiring CCU admission, congenital heart disease, idiopathic of hypertrophic cardiomyopathy, resuscitation, known or suspected thrombotic, inflammatory, or neoplastic conditions likely to be associated with an acute phase response and pregnancy.

Diagnostic procedures
Attending physicians were residents in training for cardiology, supervised by qualified cardiologists. At admission, after history, physical examination and 12 lead electrocardiogram, the attending physician estimated the probability that the patient would suffer a cardiac event within the next 6 months. This probability was marked on a double logistic probability scale.4 At discharge, 12 hours after onset of pain, the attending physician provided a new estimate for the probability that the patient would suffer a cardiac event in the 6 months to come. Between the first and second probability estimate the following diagnostic information had become available: results of blood tests (CK-MB, BSE, kidney function, glucose, cholesterol) any recurrent or resistant pain episodes and additional ECG’s.

All patients underwent Dobutamine Stress Echocardiography (DSE) within 24 hour after admission according a standard protocol.3 The physician reading the
DSE did not receive any additional patient information or test result. The DSE results were not revealed to any of the attending physicians during admission or follow-up.

Follow-up data on cardiac events were collected at 4 weeks after discharge, at 3 months and at 6 months. A cardiac event was defined as cardiac death (defined by either clinical assessment, cardiac iso-enzymes, 12 lead ECG, or autopsy), non-fatal AMI (WHO criteria), hospital admission for unstable angina and coronary revascularization procedure (PTCA or CABG), and development of new arrhythmia.

Analysis

Calibration graphs were drawn from the physicians probability estimates assessed at admission and discharge. If estimated probabilities are well calibrated they will equal the observed cardiac event rate. To assess calibration, the probability estimates were ranked and assigned to four groups, based on quartiles. Then the average probability as assigned by the physicians in each of the four groups was plotted against the proportion of patients with a cardiac event.

The result from follow-up (the reference standard) and DSE were organized in a 2x2 contingency table. Subsequently the likelihood ratio of a positive DSE (LR+) and negative DSE (LR-) with their 95% confidence intervals were estimated. The LR of a test result indicates the relative frequency of that test result in patients with a cardiac event during the 6 month follow-up period relative to the frequency of that same result in those without a cardiac event.

To evaluate the additional value of the DSE we used the likelihood ratios to estimate patient-specific post-test probabilities of a cardiac event occurring in the 6 months, using Bayes' Theorem. The probability estimate at discharge was used as the pre-test probability. Subsequently, a ROC curves was constructed using these post-test probability estimates and compared with the ROC curves of the physicians estimates at admission and discharge. Areas under the ROC curve were compared using the method of Hanley and McNeill. ROC curves can be estimated from the sensitivity (y-axes) and 1-specificity (x-axes) at each possible probability cut-off. The area under the ROC curve (AUC), a measure of discriminating performance, is 1 for perfect tests and 0.5 for tests that do not discriminate at all. The statistical software package SPSS version 11 was used for computations.

Results

A total of 611 patients were admitted, of which 203 were excluded (figure 1). Probability estimates were missing for 10 patients, so all analyses are based on 398 patients. The baseline characteristics of the 408 included patients and the 398
patients are shown in Table 1. During the 6 month follow-up period, 30 patients (7.5%) had at least one cardiac event. Table 2 provides details on the clinical outcomes.

Table 1 Clinical and demographical characteristics

<table>
<thead>
<tr>
<th>Included patients</th>
<th>Patients eligible for analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>237 (58%)</td>
</tr>
<tr>
<td>Mean age, years (SD)</td>
<td>57 (12)</td>
</tr>
<tr>
<td>History of CAD*</td>
<td>85 (21%)</td>
</tr>
<tr>
<td>Diabetes Mellitus</td>
<td>40 (9.8%)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>157 (38%)</td>
</tr>
<tr>
<td>β-Blocker</td>
<td>111 (27%)</td>
</tr>
<tr>
<td>Nitrate</td>
<td>54 (13%)</td>
</tr>
<tr>
<td>Aspirin</td>
<td>129 (32%)</td>
</tr>
<tr>
<td>Calcium antagonist</td>
<td>62 (15%)</td>
</tr>
<tr>
<td>Statine</td>
<td>67 (16%)</td>
</tr>
<tr>
<td>Ace-inhibitor</td>
<td>46 (13%)</td>
</tr>
</tbody>
</table>

*) History of coronary artery disease defined as, prior acute myocardial infarction, PTCA, CABG, or coronary artery disease documented on coronary angiograms.
The added value of pre-discharge Dobutamine Stress Echocardiography in the long-term risk stratification of chest pain patients

Table 2 The specific clinical outcomes of the 30 patients with events in the 6 months follow-up per DSE result (n=398)

<table>
<thead>
<tr>
<th></th>
<th>Positive DSE n=12</th>
<th>Negative DSE n=18</th>
<th>Total N=30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiac death</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Non-fatal AMI</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Rehosp-UA*</td>
<td>3</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>PTCA</td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>CABG</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>


The probability estimates at admission showed substantial variability, with a median value of 30% and an inter-quartile range from 10% to 50%. The mean of the physicians’ probability estimates at admission was 34%, considerably higher than the 7.5% event rate. The calibration for the probability estimate at admission (figure 2) revealed a substantial overestimation of the risk of a cardiac event: for all 4 quartiles the actual event rate was lower than the estimated risk. Probability estimates were lower when patients were discharged from the hospital but still showed variability, with a median 10% (IQR: 5% to 30%). The mean was 21%, still higher than the observed event rate of 7.5%. The adjustment to lower probability estimates at discharge is seen in the calibration graph (figure 2). It shows a consistent overestimation of the event rate in all groups. The average probability in the highest quartile was 57%, versus an observed event rate of 24%.
The DSE test was positive in 31 patients, of which 12 developed a cardiac event and negative in 367 patients, of which 349 did not develop a cardiac event. The likelihood ratio of the positive and negative DSE test result were, respectively, 7.8 (95% CI: 4.2 to 14) and 0.63 (0.47 to 1.8), see table 3.

Table 3 Measures of performance of the DSE in predicting a cardiac event (n=398, event rate 7.5%)

<table>
<thead>
<tr>
<th></th>
<th>%</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>40</td>
<td>22 to 58</td>
</tr>
<tr>
<td>Specificity</td>
<td>95</td>
<td>93 to 97</td>
</tr>
<tr>
<td>Likelihood ratio pos</td>
<td>7.8</td>
<td>4.2 to 14</td>
</tr>
<tr>
<td>Likelihood ratio neg</td>
<td>0.63</td>
<td>0.47 to 1.8</td>
</tr>
</tbody>
</table>

The performance of the physicians in identifying patients at high risk for a cardiac event in the 6 months can be assessed by inspection of the receiver operating characteristic (ROC) curves (figure 3), and the area under them (AUC). The AUC of the physicians’ probability estimates at admission was 0.76 (95% CI: 0.68 to 0.84), not statistically significant better than the AUC of 0.83 (95% CI: 0.77 to 0.90) at discharge (p=0.07). The AUC for the post-test probabilities after the DSE (AUC=0.87, 95% CI: 0.81 to 0.93) was not significantly better compared to the physicians’ probabilities at discharge (p=0.11). (see figure 3).

We looked at the value of the DSE test in specific subgroups of patients (table 4). The group of patients in the lowest quartile, with an estimated probability at discharge of 0 to 5%, had a zero event rate. Even after a positive DSE, the risk of
6 The added value of pre-discharge Dobutamine Stress Echocardiography in the long-term risk stratification of chest pain patients

an event will be extremely low. In this low risk group (145/408 patients), the DSE has no additional value in risk assessment. The upper quartile consist of 87 patients, with an estimated probability of 31 to 100%. After a negative DSE the probability of an event is still 16%. In this subgroup at high risk the DSE has limited additional value in risk assessment either: all patients remain at elevated risk, even after a negative DSE result. In the remaining group, the two intermediate quartiles consisting of 166 patients with an estimated probability between 6% and 30%, the cardiac event rate is 5.4%. In this subgroup (166/408), the event rate for patients with a positive test result is 25%, whereas the event rate for a negative test is 3.9%, much lower than the overall event rate of 7.5%.

Table 4 Subgroups of patients, based on the probability estimate at discharge and the event rate per group.

<table>
<thead>
<tr>
<th>Estimated probability range</th>
<th>Cardiac Event rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSE +</td>
</tr>
<tr>
<td>Low risk 0 to 5%</td>
<td>0%</td>
</tr>
<tr>
<td>Intermediate risk 6 to 30%</td>
<td>25%</td>
</tr>
<tr>
<td>High risk 31 to 100%</td>
<td>64%</td>
</tr>
</tbody>
</table>

Discussion

We evaluated the added value of the pre-discharge DSE in the long-term risk stratification of chest pain patients with non-diagnostic or normal ECG and with ruled out acute coronary syndrome at the emergency room. In this population the overall event rate in the follow-up was 7.5%. The calibration of the physicians in estimating the probability of an event was modest, with overestimated cardiac event rates. This tendency to overestimate is typical for physicians in general. In contrast, the discriminating performance of the physicians can be considered as good, with an AUC of 0.83. The area under the curve of the post-DSE probabilities only showed a minor improvement relative to the AUC of the physicians, indicating a modest added value of the DSE. It seems that only in a subgroup of patients at intermediate estimated risk, consisting of less than half of the patients in the study group, the DSE is able to differentiate between clinical relevant high risk and low risk patients.

One potential limitation of this study is the absence of post-test probabilities as estimated by the physicians themselves, incorporating the results of the DSE. A direct comparison of such a post-test probability estimate with the probability estimate at discharge was thus not possible. This was inherent in the study design,
which called for a total blind evaluation of the DSE, to prevent that a positive test result would be incorporated in the decision to perform an intervention defined as one of the endpoints. Instead we calculated posttest probabilities using Bayes Theorem, using a single set of likelihood ratios for all patients, based on the observed event rates.

A number of other studies have looked at stress echocardiography in admitted chest pain patients with negative or non-diagnostic electrocardiograms, reporting promising results of the DSE.1 2 10 We are not aware of literature on the added value of DSE in these particular patients.

We've previously published results of the prognostic value of the DSE in predicting a cardiac event in this study population.3 By means of multivariable logistic regression, the DSE was found to carry independent prognostic value, with an odds ratio of 10.7 in troponin negative patients. Despite the statistically significant association, the present study shows that the added clinical value may be limited to patients in whom the attending physician attributes an intermediate risk of subsequent cardiac events.

The less than perfect sensitivity and high specificity of the DSE result in a poorly informative likelihood ratio of a negative test result. In other words, a negative DSE test is unable to exclude future events.11 In this study, 92% of the patients had a negative DSE test. In addition the event rate was rather low, 7.5%. Tests in general tend to perform poorly in populations with a low event rate.12

We sought to identify subgroups in which the DSE may have additional discriminating value. We found that patients at low risk (0 to 5%), as estimated by the physician at discharge, can be discharged safely without performing a DSE test. The test will most likely be negative, and the event rate after a positive test result is still very low. Alternatively, in the patients where the physicians assigned a high probability to (30% or more), the probability of an cardiac event is elevated, even after a negative DSE result. Only in an intermediate group (estimates risks of 6-30%) the DSE can be used to distinguish patients at relatively low risk from patients at high risk, although the event rate will still be nonzero, even after a negative result. Naturally, it may be important to use a test to lower the risk from 7.5% to 3.9%.

Despite a well-established prognostic association the, added value of the pre-discharge DSE is modest in the risk stratification of patients with chest pain and a non-diagnostic electrocardiogram, as physicians perform rather well in the risk stratification of these patients. Other tools will be needed to improve risk assessment further.
6 The added value of pre-discharge Dobutamine Stress Echocardiography in the long-term risk stratification of chest pain patients

References
Hoofdstuk titel
Vervolg hoofdstuk titel
Abstract

Background - Ankle decision rules help to determine which patients with ankle injuries should undergo radiography. However, these rules are limited by imperfect generalizability and sensitivity. The judgement of physicians, aided by structured data collection, is a potential alternative. We compared the diagnostic performance of 2 decision rules with the performance of physicians, aided by structured data collection, in ruling out fracture in patients with acute ankle injury. Methods - Consecutive patients with acute ankle injury who visited the emergency department of a teaching community hospital in Amsterdam were included in the study. After taking the patient's history and performing a physical examination, the surgical resident in each case completed a specially developed structured data form incorporating all of the variables in the Ottawa and Leiden ankle rules, as well as some additional variables. The form then asked whether the resident thought radiography was necessary. Each patient then underwent ankle and midfoot radiography. The films were independently interpreted by a radiologist and a trauma surgeon, who were both blinded to the information on the data form. Sensitivity, specificity and the percentage of patients for whom radiography was recommended were the main outcome measures.
7 Physicians’ performance using structured data collection versus diagnostic decision rules in acute ankle injury

Results - Of 690 consecutive patients, 647 met the inclusion criteria. Fractures were observed in 74 (11%) of these patients. Sensitivity was 89% (95% confidence interval [CI] 80% to 95%) for the Ottawa ankle rules, 80% (95% CI 69% to 88%) for the Leiden ankle rule and 82% (95% CI 72% to 90%) for physicians’ judgement. Specificity was 26% (95% CI 23% to 30%), 59% (95% CI 55% to 63%) and 68% (95% CI 64% to 71%) respectively. Radiography was recommended in 76% (95% CI 72% to 79%), 46% (95% CI 42% to 50%) and 38% (95% CI 34% to 42%) of cases respectively. The Ottawa rules missed 8 fractures, of which 1 was clinically significant, the Leiden rule missed 15 fractures, of which 5 were clinically significant, and the residents missed 13 fractures, of which 1 was clinically significant.

Interpretation - Physicians’ judgement, aided by structured data collection, was similar to existing international and local decision rules in terms of sensitivity in identifying cases requiring radiography and may outperform these prediction rules in terms of minimizing radiographic examinations for patients with ankle trauma.

Introduction
It is not uncommon for physicians to routinely order radiography for patients with ankle injury, although less than 15% of such patients actually have fractures. This policy is safe, in that it ensures that no fractures are missed, but it entails a high use of resources. Stiell and colleagues suggested that guidelines might help physicians to identify patients who did not have fractures. In these cases radiography could be safely withheld and the associated costs avoided.

Empirical evidence for the claim that decision rules perform better than physicians’ judgement came from 2 studies by Stiell and colleagues, both published in 1992, which used a set of rules now known as the Ottawa ankle rules. Shortly afterward, Stiell and colleagues expanded the Ottawa ankle rules. Some hospitals have constructed and implemented locally developed decision rules, whereas others have modified the Ottawa rules to suit their particular systems.

The Ottawa ankle rules are among the most validated and most widely implemented clinical prediction rules. Yet they cannot escape from imperfect generalizability (the ability to perform as well in other populations as in the population for which they were originally developed), a common problem with prediction rules. The Ottawa rules have proved unsuccessful in some populations. A possible explanation for this problem might be more or less severe ankle injury in different populations because of various thresholds for seeking medical assistance. Conversely, it might be questioned whether physicians fare so poorly in making recommendations about radiography for patients with ankle injury that formal
decision rules and explicit calculations are needed. There is evidence that structured data collection (completion of a form with various questions relevant to a specific decision process and an explicit question about the outcome of the decision) might also improve the diagnostic performance of physicians.\textsuperscript{13,15} Such a form might be an alternative to formal decision rules.

We hypothesized that even relatively inexperienced surgical residents in the emergency department can make accurate judgements in one of the most common situations encountered in this setting. In this prospective study, we compared the diagnostic performance of physicians aided by structured data collection with the performance of the Ottawa ankle rules and a local decision rule (the Leiden ankle rule) for patients with acute ankle injury.

Methods
This prospective comparative study was performed at a medium-sized teaching community hospital in Amsterdam, the Netherlands. Consecutive patients presenting to the emergency department with acute ankle injury between January 1998 and April 1999 were eligible. Acute ankle injury was defined as any case of painful ankle resulting from trauma. The ankle was defined as the malleolar area and the midfoot area, both of which are commonly involved in twisting injuries. Patients were excluded if they were under 18 years of age, if they were pregnant, if they had been referred with radiographs from another hospital or a general practitioner, if the ankle injury had occurred more than 5 days previously, if they had returned for reassessment or if they had experienced multiple trauma. No changes in clinical management were made as a result of this study, so approval was not sought from the local ethics committee, nor were the patients asked to provide informed consent.

Two sets of decision rules for ankle injury were used. The Ottawa ankle rules were developed in 1992\textsuperscript{1} and refined in 1993.\textsuperscript{4} These rules consist of a foot section and an ankle section, each comprising 3 unweighted variables. If any one of these variables is positive, radiography is indicated (Table 1). The rules were designed to have 100% sensitivity in detecting clinically significant fractures.

The Leiden ankle rule was developed at the university hospital of the city of Leiden in 1991.\textsuperscript{6} It consists of 7 rows of which each row consists of one or more variables. If per row at least one variable is positive, the stated score (weighted score) is given to that row, except for the last row, which depends on the age of the patient. If two variables are positive, the score is not doubled. For example, if both deformity and crepitation are positive, the score for that row is 5. The final score is the sum of the row scores (Table 2). The developers of the rule reported a
Physicians’ performance using structured data collection versus diagnostic decision rules in acute ankle injury

sensitivity of 100% in detecting clinically significant fractures at this cutoff level. The attending physicians in this study were junior surgical and orthopedic residents. For each case, the resident was asked to complete a structured data collection form, after taking the history and performing a physical examination. The one-page form, developed specifically for this study, incorporated the criteria for the Ottawa ankle rules and the Leiden ankle rule, along with some additional variables (Figure 1, translated from Dutch). At the end of the form, the resident was asked to indicate whether radiographic examination was necessary. In addition, the resident was asked to estimate the likelihood of fracture on a probability scale ranging from 0% to 100%. The residents were instructed in scoring the various items on the form. They were told that the study involved a comparison between the Ottawa ankle rules and a local decision rule, but the rules as such were not discussed. Although some residents may have been aware of the rules, it was assumed that they did not use any specific rules in making the decision to request radiography or in estimating the probability of fracture.

After the resident had completed the form, the patient underwent a radiographic series of both foot and ankle, regardless of the resident’s assessment of the need for radiography. A radiologist and trauma surgeon (JNK) interpreted the radiographs independently. Both were blinded as to the contents of the structured data collection form and any treatment given. Disagreement was resolved by consensus. A fracture was defined as any fresh fracture line. A clinically significant fracture was defined as an apparent dislocation of more than 2 mm and a fracture line more than 3 mm across. Patients with clinically significant fractures, as diagnosed by the resident, received operative treatment or cast immobilization.

For each patient, the scores for the Ottawa and Leiden ankle rules were calculated from the relevant variables on the structured data collection form. The radiographic series was used as the reference standard. For both the Ottawa and the Leiden ankle rules, we calculated the sensitivity, specificity, percentage of missed fractures and percentage of patients for whom radiography would be indicated, on the basis of the established cutoff scores for these rules. The percentage of patients for whom radiography would be indicated was calculated as the percentage of true positives and false positives for the whole patient group.

We calculated the same items as determined by residents’ judgement; for the residents, the percentage of patients for whom radiography would be recommended was based on answers to the question of whether radiographic examination was deemed necessary (see Figure 1).

Receiver operating characteristic (ROC) curves were created for the 2 sets of decision rules. To draw the ROC curve for the Ottawa rules, sensitivity and
specificity were calculated on the basis of combined foot and ankle criteria for the following 5 thresholds: 0 items positive through 5 items positive. The area under each ROC curve (AUC) was subsequently calculated. The AUC expresses the performance of a diagnostic tool in distinguishing patients with the target condition from those without it for all possible cutoff values.

We also calculated the AUC for the residents’ estimates of the probability of fracture. A calibration curve was constructed for the probability estimates to examine the reliability of the residents’ predictions. For this purpose the probability estimates were divided into deciles. For each decile the actual fracture rate was calculated and compared with the mean probability estimate for all patients in that decile. With well-calibrated probability estimates, the average probability estimate should correspond to the actual fracture rate in each decile.

The McNemar test for paired samples was used to test for significant differences between the rules and the residents in terms of the percentage of patients for whom radiography was recommended. The method of Hanley and McNeil was used to test whether differences between the AUCs were statistically significant. Confidence intervals (CI) around estimates were calculated when appropriate.

The sample size needed for this study was estimated with the McNemar test. We considered as clinically important a 5% difference between the decision rules and the residents’ answers to the yes/no question in terms of patients for whom radiography was recommended. We calculated that to achieve a power of at least 90%, with a 2-tailed 5% type 2 error, a (paired) sample size of at least 609 subjects was necessary.

Table 1 Ottawa ankle rules

<table>
<thead>
<tr>
<th>Ankle Radiographic Series Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>An ankle radiographic series is only required if there is any pain in the malleolar zone and any of these findings:</td>
</tr>
<tr>
<td>Bone tenderness at posterior edge (6 cm) or tip of lateral malleolus.</td>
</tr>
<tr>
<td>Bone tenderness at posterior edge (6 cm) or tip of medial malleolus</td>
</tr>
<tr>
<td>Inability to bear weight both immediately and in Emergency Department</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Foot Radiographic Series Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A foot radiographic series is only required if there is any pain in the midfoot zone and any of these findings:</td>
</tr>
<tr>
<td>Bone tenderness at base of fifth metatarsal</td>
</tr>
<tr>
<td>Bone tenderness at navicular bone</td>
</tr>
<tr>
<td>Inability to bear weight both immediately and in Emergency Department</td>
</tr>
</tbody>
</table>
Physicians’ performance using structured data collection versus diagnostic decision rules in acute ankle injury

Table 2 Leiden rule.

<table>
<thead>
<tr>
<th>Clinical Picture</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deformity / Instability / Crepitating</td>
<td>5</td>
</tr>
<tr>
<td>Weight bearing</td>
<td>3</td>
</tr>
<tr>
<td>Pulseless / weakened posterior tibial artery</td>
<td>2</td>
</tr>
<tr>
<td>Pain on palpation malleoli / Metatarsal V</td>
<td>2</td>
</tr>
<tr>
<td>Swelling malleoli / Metatarsal V</td>
<td>2</td>
</tr>
<tr>
<td>Swelling / pain Achilles tendon</td>
<td>1</td>
</tr>
<tr>
<td>Age divided by 10:</td>
<td>...</td>
</tr>
<tr>
<td>Total &gt; 7: radiographs</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1 Data collection sheet with variables from history and physical examination
Paleness: a noticeable difference with the contra lateral side.
Traffic accident: if trauma was caused by a traffic accident.
Pulseless or weakened posterior tibial artery: a marked difference with the contra lateral side. Axial Compression: pain when axial pressure is applied (pressure plantar on the heel in direction towards the knee).

ANKLE INJURY DATA COLLECTION FORM

Patient information like patient ID number, sex and age

<table>
<thead>
<tr>
<th>Inspection</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unable to bear weight (4 steps)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deformity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hematoma</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paleness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swelling Malleol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swelling MT V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swelling Achilles tendon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic accident</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Examination</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain on palpation lat. malleolus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain on palpation med. malleolus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain on palpation back MT V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain on palpation os navicular</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain on palpation Achilles tendon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pulseless a. tibialis post.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crepitation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Axial compression pain</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What do you think is the probability of ankle fracture in this patient after considering the information from history and physical examination?

What do you want to send this patient for radiographic examination?

Name physician: _______________________
Function: _______________________

102
Results
Twenty-four residents gathered data during the study period. A total of 690 patients presented with acute ankle trauma, of whom 647 were included in the study (Figure 2). Mean age was 35 (range 18 to 92, standard deviation 14) years. Half of the subjects, 324, were female. Fractures were observed radiographically for 74 (11%) of the patients, and 41 of these were considered clinically significant. Nineteen of the patients underwent operative treatment, and the other 22 underwent cast immobilization. The 33 patients with insignificant fractures and all patients without a fracture were treated with tape or bandage.

The Ottawa ankle rules identified 66 of the 74 fractures (sensitivity 89% and specificity 26%), the Leiden ankle rule identified 59 (sensitivity 80% and specificity 59%), and the residents identified 61 (sensitivity 82% and specificity 68%) (Table 3). The Ottawa rules missed 8 fractures, of which one was clinically significant (a Weber B fracture of the fibula, which was treated with a cast). The Leiden rule missed 15 fractures, of which 5 were clinically significant (2 metatarsal fractures,
Physicians’ performance using structured data collection versus diagnostic decision rules in acute ankle injury

1 anterior fracture of the calcaneus, 1 Weber A fracture and 1 Weber B fracture, all of which were treated with a cast). The residents missed 13 fractures, of which 1 was clinically significant (a Weber A fracture with dislocation that needed repositioning and cast treatment). The Ottawa rules and the Leiden rule recommended radiography in 76% and 46% of the cases, respectively. The residents considered radiography necessary in 38% of the cases.

Table 3 Diagnostic performance indicators of the ankle rules and physician’s performance in distinguishing fracture from non-fracture

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity % (CI)</th>
<th>Specificity % (CI)</th>
<th>AUC % (CI)</th>
<th>X-rays % (CI)</th>
<th>Missed clinical significant fractures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ottawa</td>
<td>89 (80-95)</td>
<td>26 (23-30)</td>
<td>69 (62-76)</td>
<td>76 (72-79)</td>
<td>1</td>
</tr>
<tr>
<td>Leiden</td>
<td>80 (69-88)</td>
<td>57 (55-63)</td>
<td>77 (71-83)</td>
<td>46 (42-50)</td>
<td>5</td>
</tr>
<tr>
<td>Physician†</td>
<td>82 (72-90)</td>
<td>68 (64-71)</td>
<td>80 (74-87)</td>
<td>38 (34-42)</td>
<td>1</td>
</tr>
</tbody>
</table>

*) X-rays: percentage of patients, which should receive radiographic assessment according to the results of the decision rules or the physicians decision. (calculated as (true positives + false positives)/total x 100%)
†) The AUC of the Ottawa rules was significant different from both the Leiden rule and the probability estimate of the physician (Hanley and McNeil test. p<0.03 and p<0.01 respectively).
‡) The sensitivity, specificity and X-rays are based on the yes/no question, the AUC is based on the probability estimates.

Figure 3 Receiver operating characteristic curves of the ankle rules and the probability estimates of the physicians.

- Ottawa, - Leiden, - Physicians
McNemar tests showed a significant difference between both sets of decision rules and the residents’ opinions in terms of the percentage of cases in which radiography was recommended (p < 0.001 for all 2-way comparisons). The ROC curves are shown in Fig. 3. The AUC for the Ottawa rules was 0.69, significantly lower than the AUC for both the Leiden rule (0.77) and the residents’ probability estimates (0.80) (p < 0.05). There were no significant differences between the Leiden rule and residents’ probability estimates.

The calibration curve shows 8 point estimates, covering a total of 10 deciles (Fig. 4). The first point, at zero, comprises 3 deciles representing 243 patients, for all of whom the estimated probability of fracture was zero. The observed proportion of fractures in these patients was 4.1%. The calibration curve of the residents’ probability estimates was nearly perfect for the low probability estimates, indicated by the fact that estimates in the range from zero to about 10% are close to the dashed line. Above 10% estimated probability the calibration curve shows increasing overestimation: in the highest decile, the mean estimated probability of fracture was 85%, which contrasts sharply with the 65% observed fracture rate.

**Interpretation**

In this study the diagnostic performance of residents using structured data collection was compared with 2 sets of diagnostic decision rules for patients with acute ankle injury. These rules were designed for determining whether radiographic assessment is necessary to exclude fracture. The sensitivity of diagnostic performance by relatively inexperienced surgical and orthopedic residents, after structured data collection, was similar to the sensitivity for the Ottawa rules and the local Leiden rule. In addition, the residents outperformed both sets of rules in terms of minimizing the percentage of patients for whom radiography was deemed necessary.
The results of this study may have been biased by a Hawthorne effect, whereby participants (in this case, the residents examining the patients) tend to perform better simply because of their awareness that they are participating in a comparative study. We feel that the effects of the Hawthorne effect in this study were probably not large, in that the residents were informed that the study involved a comparison between decision rules, not an assessment of their own performance.

To rule out verification bias, every consecutive patient with ankle injury underwent standard radiography of both ankle and foot. This could have effected the residents’ decisions, because they knew in advance that radiographs would be obtained regardless of their clinical judgement. The residents were probably less careful than they would have been in a nonstudy situation. This form of bias might have led to more false-negative and fewer false-positive judgements and thus to underestimation of sensitivity and overestimation of the potential reduction in radiographic assessment.

We instructed the residents to focus on distinguishing fractures from nonfracture injury; there was no focus on potential treatment. The residents may have based their clinical judgement on possible treatment options and, consequently, shifted their diagnostic frame of reference to more severe cases. For example, if a significant fracture could be ruled out, radiography would have been deemed unnecessary since the treatment for a small fracture is the same for no fracture. Such a shift in the frame of reference might lead to an underestimation of sensitivity due to missing small fractures.

To our knowledge only 4 studies on the use of structured data collection are available, all focusing on the diagnosis of acute abdominal pain. These studies reported better diagnostic performance by physicians who used structured data collection compared to standard policy that was not aided by decision-making tools. Structured data collection enables consistent clinical assessment and as such might induce a better thought-out decision, which would be particularly beneficial for junior (less experienced) residents.

A small number of previous studies have compared ankle rules or guidelines with physicians’ performance, but none of these studies used explicit structured data collection. Stiell and colleagues have published 2 studies on this subject. The first concerned estimates of the probability of clinically significant fractures. The authors concluded that if the threshold for requesting radiographs was less than 10%, the physicians performed reasonably well, missing only 2 of 145 significant fractures (sensitivity 99% [95% CI 95% to 100%]). We used the probability estimate as an intermediate measure, to conceptualize the decision process of the physician without dichotomizing it. Nevertheless, applying a threshold of 10% for
the data in this study would have yielded a sensitivity of 90% (95% CI 77% to 97%) for detecting significant fractures. In the study by Stiell and colleagues' no ROC analysis was performed to compare overall diagnostic performance. A second publication by the same authors focused mainly on physicians' performance in detecting clinically significant fractures. For comparative purposes, we created the ROC curve for clinically significant fractures only in our study; the AUC was 0.89, almost equal to the AUC of 0.88 for the probability estimates reported by Stiell and colleagues.

In our population the 2 sets of ankle rules and the residents missed some fractures. Some might argue that it is unethical to allow any fractures to be missed, even if they are small and quality of care is not affected. However, both sets of rules and the residents missed clinically important fractures, which would affect quality of care. A possible explanation for the lack of perfect sensitivity for the Ottawa rules in this study may be the limited experience of the physicians. Most of them had just finished their medical training and were much less experienced than the well-trained physicians in the studies of Stiell and colleagues. Unknown population differences might also be a factor, although there were no obvious reasons why the populations would not be comparable.

Another disadvantage of prediction rules is the lack of transportability, the ability to use the rules in related populations. For example, both the Ottawa and Leiden rules do not apply to patients with specific comorbidity, such as neuropathic disorders. We believe that procedures that continuously challenge the decision-making skills of physicians are in the long run more accurate than rigid guidelines.

We have shown that the clinical skills of relatively inexperienced physicians, supported by structured data collection and explicit questioning, allowed them to perform as well as validated decision rules in judging the need for radiography in patients with acute ankle injuries. Explicit ankle rules are not necessary to make a decision about the need for radiography to rule out fracture in a patient with a painful ankle after trauma. Even more may be at stake: these results indicate that even greater reductions in the use of radiography can be attained, without compromising quality of care, by relying on physicians' judgment rather than on formalized decision rules.
7 Physicians’ performance using structured data collection versus diagnostic decision rules in acute ankle injury

References


Abstract

Introduction - It is an empirical question to what extent physicians adjust their initial uncertainty after diagnostic information becomes available. The aim of this study is to investigate the interpretation of V/Q lung scan results by physicians in the diagnosis of pulmonary embolism (PE).

Methods - Data were obtained in a prospective multi-center study in patients with suspected PE. Treating physicians were asked to assign the clinical probability of PE before and after V/Q scans were performed. This made it possible to calculate the subjective likelihood ratio (LR) of scan results. The objective LR was calculated from a blinded comparison of test results with a conjoint reference standard (independent interpretation of V/Q scan and pulmonary angiography). Calibration curves were estimated using the clinical probability estimates.

Results - The mean subjective LR of a normal V/Q scan result was 0.056 (95% CI: 0.038 to 0.082) compared to an objective LR of 0.046 for the same result (95% CI: 0.012 to 0.19). These values were 26 (CI: 16 to 41) versus 20 (CI: 8.8 to 43) for a high probability scan and 1.1 (CI: 0.77 to 1.5) versus 0.43 (CI: 0.31 to 0.62) for a non-diagnostic scan. No significant center differences were found for non-diagnostic scan results, although centers more or less adjust for center differences in the diagnostic worth of a non-diagnostic scan result. Overall calibration was good.

Conclusion - In general physicians are well calibrated and they update prior probabilities according to the inherent information in a normal and high-probability lung scan result. A non-diagnostic scan is interpreted inconsistent to its diagnostic worth.
Introduction
The diagnosis of pulmonary embolism represents a clinical challenge, due to the limited specificity of presenting signs and symptoms and the less-than-perfect accuracy of additional diagnostic tests. A number of research efforts have aimed at the development of safe and cost-effective diagnostic strategies for suspected pulmonary embolism. In many of these strategies, the ventilation-perfusion (V/Q) lung scan plays a crucial role, despite its limited overall sensitivity and specificity. The hitch of the lung scan lies in the non-diagnostic test result, obtained in about half of the patients suspected for pulmonary embolism, of which 10 to 30% actually have the disease when subjected to further testing.

The limitations of all imaging modalities for pulmonary embolism - lung scanning, spiral computerized tomography, pulmonary angiography and venous leg imaging - have necessitated the use of these tests in algorithms and in combination with clinical pretest probability assessment in diagnostic management algorithms. Rodger and Wells considered clinicians to be adept at assigning such pretest probabilities after overall clinical assessment, and incorporating them in the further workup of pulmonary embolism suspected patients. The clinical strength of suspicion after clinical assessment will vary, depending on the patient's complaints, risk factors, and the likelihood of other conditions. Yet pretest probabilities are susceptible to error. Multiple formal clinical models, explicit prediction and decision rules have been developed to improve the accuracy of pretest probability assessment.

Expressing one's clinical strength of suspicion in probability form can be done both before and after testing. Bayes' theorem puts a formal link between these two probabilities: the posttest odds should equal the pretest odds for pulmonary embolism multiplied by the likelihood ratio of the test result. The likelihood ratio of a test result expresses how much more likely that test result is in patients with pulmonary embolism compared to those without it.

It remains an empirical question how well physicians adjust their uncertainty based on the information in the test result. Highly informative test results, with likelihood ratios that are either close to zero or far above unity, should lead to a more pronounced reduction in uncertainty compared to test results with likelihood ratios closer to unity. In addition, the reduction in uncertainty should be proportional to the strength of the prior suspicion. Imperfect pre-test probabilities and incorrect likelihood ratios can lead to flawed post-test probabilities and, subsequently, to faulty therapy decisions.

The purpose of this study was to investigate the judgment of physicians in interpreting lung scan results in the diagnosis of pulmonary embolism, to see to
what degree they update prior probabilities. A consecutive series of patients with a clinical suspicion for pulmonary embolism underwent a standard diagnostic algorithm, starting with V/Q scanning and followed, if necessary, by further testing. Physicians could express pre-test and post-test probabilities of pulmonary embolism on a visual analogue scale, immediately before and after lung scanning. The calibration of respective probabilities were evaluated and the physician's subjective likelihood ratios - obtained from the pre and posttest probabilities - were compared with the objective likelihood ratios of the V/Q lung scan - obtained by comparing the lung scan results with the final results of the reference standard.

Patients and methods

Patients
Data were obtained as part of a prospective study in six Dutch teaching hospitals, reported in detail elsewhere. From May 1997 through March 1998, consecutive in- and outpatients with a clinical suspicion of pulmonary embolism (PE) were eligible for the study. Patients were excluded if they were younger than 18 years of age, were pregnant, had already undergone objective diagnostic tests for PE, had an indication for acute thrombolytic therapy, or if there was an expected inability to complete the study protocol within 48 hours of presentation. The Institutional Review Boards of all participating hospitals had approved the study protocol. Informed and consenting patients were included in the study.

Diagnostic Investigations
Within 24 hours of study inclusion a detailed medical history, compression ultrasonography and a ventilation-perfusion (V/Q) lung scan were performed in all patients. Spiral CT was only performed in patients with an abnormal perfusion scan. Pulmonary angiography was performed in patients with a non-diagnostic lung scan and if a high probability lung scan was followed by a normal spiral CT. The maximum allowed time for completion of this diagnostic protocol was 48 hours.

A six-view perfusion lung scintigraphy was obtained after the administration of 100 MBq of 99mTc technetium-labelled macro-aggregates of albumin. If segmental or larger perfusion defects were seen, ventilation lung scintigraphy was added using 81mKr Krypton gas. In each center, experienced nuclear medicine physicians interpreted the lung scans by using a lung segment reference chart. Lung scans were classified according to previously described criteria as either normal (no perfusion defects), high probability (at least one segmental or larger perfusion
defect with locally normal ventilation) or non-diagnostic (ventilation-perfusion abnormalities not meeting the criteria for normal or high probability) (12, 13). Lung scans of poor quality due to technical failure were classified as non-interpretable.

Pulmonary angiography was performed using a digital subtraction technique with the catheter positioned selectively in the right and left pulmonary artery and was interpreted independently by two radiologists according to accepted criteria. If there was no consensus, the independent interpretation of a third reader was considered decisive.

A central panel re-interpreted independently all lung scans and pulmonary angiograms, without knowledge of the outcomes of ultrasonography, locally interpreted V/Q scan or, if applicable, the spiral CT. The panel's conclusion was used as the reference standard. Patients were classified as having PE in case of a high probability lung scan or abnormal angiogram, while the diagnosis was refuted if the lung scan or angiogram was normal. Only those patients in whom the diagnosis of PE was confirmed or refuted per-protocol were included in the analysis.

Probability estimates
Upon presentation of a patient and prior to all other diagnostic investigations, the treating physician was asked to assign a clinical probability of pulmonary embolism using a non-linear, logistic visual analog scale, ranging from 0 to 100 percent. This estimate could solely be based on symptoms and signs and results of routine tests (arterial blood gas analysis, electrocardiogram and chest X-ray), if available. A second scale, identical to the first, was used in assigning a post-test clinical probability, after the results of the locally interpreted V/Q scan had become available. The probability estimates were not communicated to the central panel.

Data analysis
If subjective probabilities are well calibrated the assigned probability of PE should equal the observed proportion of patients with PE. For example, of all patients assigned a 10% probability of PE, 10% is expected to harbor a pulmonary embolism.

To study calibration, the probability estimates were ranked and assigned to ten groups, based on deciles. Calibration curves were drawn for the pre-test and post-test probabilities, plotting the average probability as assigned by the physicians in each of the ten groups against the proportion of patients with a pulmonary embolism. Likelihood ratios of the results of the V/Q scan were estimated by comparing
the results of the locally interpreted V/Q scan with the results from the conjoint reference standard, as assigned by the panel. For the purpose of this paper, these likelihood ratios will be referred to as objective likelihood ratios. 95% Confidence intervals for the objective likelihood ratio were calculated using Simel’s method.16

For each pair of a pre-test and post-test probability, assigned to the same patient, a subjective likelihood ratio was calculated by transforming the pre-test and post-test probability into odds, and taking the ratio of post-test odds against pre-test odds.17 To avoid undefined results, probabilities of 0 and 100 were replaced by probabilities of 0.001 and 99.999 respectively. Since the estimated subjective likelihood ratios do not follow a normal distribution, a logarithmic transformation was performed. The average subjective likelihood ratio of each test result of the V/Q scan was then calculated as the back-transformed (geometric) mean of the logarithmically transformed subjective likelihood ratios. Confidence intervals (CI) were calculated on a log-scale as mean lnLR ±1.96*SE_{mean} and then back-transformed to the original scale.

To explore whether physicians modified for center differences in (objective) likelihood ratio, we compared them with the subjective likelihood ratio per center. All computations were done with the statistical software package SPSS version 10.

---

<table>
<thead>
<tr>
<th>Patients eligible for inclusion (n=1162)</th>
<th>Excluded Patients (n=179)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients eligible for consent (n=983)</td>
<td></td>
</tr>
<tr>
<td>No informed consent (n=356)</td>
<td></td>
</tr>
<tr>
<td>Informed consent (n=627)</td>
<td></td>
</tr>
<tr>
<td>No final diagnosis (n=110)</td>
<td></td>
</tr>
<tr>
<td>Final diagnosis (n=517)</td>
<td></td>
</tr>
<tr>
<td>Pre- and post test probability available (n=275)</td>
<td>Pre- and/or post test probability not available (n=242)*</td>
</tr>
</tbody>
</table>

* Patients were excluded for the following reasons: protocol expected to be >48 hours n=104, prior diagnostic testing n=43, age <18 years n=16, pregnancy n=11, indication for acute thrombolytic therapy n=5. 1) Reasons pre-test probability not available: physician unreachable n=31, result lung scan already available n=27, physician refusal n=7, not specified n=39. Reasons post-test probability not available: physician unreachable n=11, result angiography available n=2, not specified n=19, not applicable n=106.

Figure 1. Patient flow-chart.

---
8 Physicians’ interpretation of ventilation-perfusion lung scans

Results
A total of 1162 consecutive patients with clinically suspected pulmonary embolism were screened, of whom 179 had to be excluded on the basis of one of the predefined criteria. (See flowchart, figure 1.) Of the 983 eligible patients, 627 (64%) gave written informed consent. A reference diagnosis regarding the absence or presence of pulmonary embolism could not be reached in 110 of these 627 patients, due to withdrawal of informed consent, clear evidence for an alternative diagnosis, medical reasons, or technical failure. For 242 of the 517 remaining patients a pre-test or posttest probability was missing.

The diagnosis of pulmonary embolism was eventually confirmed in 114 patients of the remaining 275 (41%). In this group, 23% had a normal scan (n=63), 43% a non-diagnostic scan (n=119), 32% a high probability scan (n=89) and 1.5% a non-interpretable scan (n=4). The prevalence of pulmonary embolism in these 4 groups was 2/63=0.032, 28/119=0.24, 83/89=0.93 and 1/4=0.25 respectively. A total of 140 different physicians were involved in assigning the probabilities, of which 10% were estimated by specialists. Table 1 presents the characteristics of the patients used in the analyses.

Table 1. Clinical and demographical characteristics.

<table>
<thead>
<tr>
<th></th>
<th>Study patients (n=627)</th>
<th>Available for the analysis of probabilities (n=275)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>270 (43%)</td>
<td>119 (43%)</td>
</tr>
<tr>
<td>Mean age, years (SD)</td>
<td>53 (18)</td>
<td>54 (18)</td>
</tr>
<tr>
<td>Outpatients</td>
<td>490 (78%)</td>
<td>212 (77%)</td>
</tr>
<tr>
<td>Median duration of symptoms, days</td>
<td>3 (1-9)</td>
<td>3 (1-8)</td>
</tr>
<tr>
<td>Symptoms of DVT†</td>
<td>43/624 (6.9%)</td>
<td>17 (6.0%)</td>
</tr>
<tr>
<td>Surgery</td>
<td>111/626 (18%)</td>
<td>46 (17%)</td>
</tr>
<tr>
<td>Trauma</td>
<td>38/623 (6.1%)</td>
<td>16 (6.0%)</td>
</tr>
<tr>
<td>Immobilization</td>
<td>100/622 (16%)</td>
<td>43 (16%)</td>
</tr>
<tr>
<td>Previous VTE‡</td>
<td>98/622 (16%)</td>
<td>46 (17%)</td>
</tr>
<tr>
<td>Family history of VTE</td>
<td>122/551 (20%)</td>
<td>58/249 (23%)</td>
</tr>
</tbody>
</table>

*) Patients of which data available for calculation of the subjective LR.
†) DVT = deep vein thrombosis
‡) VTE = venous thromboembolism
Probabilities and calibration

The pre-test probability estimates were distributed across the whole range, with an inter-quartile range (IQR) from 30% to 75%. The overall calibration for the pre-test probability was good but not perfect, as the probabilities do not tend to match the observed proportions of pulmonary embolism. The calibration curve shows that physicians tended to overestimate the risk of pulmonary embolism when they assigned higher probabilities. (Figure 2).

Figure 3 shows scatter plots of the pre-test versus post-test probability estimate for the three local lung scan result categories. Overall low post-test probabilities were assigned to patients with a normal scan (IQR: 1% to 5%) and high probabilities to patients with a high-probability scan (IQR: 87% to 100%). The probabilities after a non-diagnostic scan show far more variability, with a substantial number of patients with a very low or high post-test probability estimate of pulmonary embolism (IQR 30% to 80%). In some patients a high pre-test probability was followed by a very low post-test probability after a non-diagnostic scan. In others, the reverse could be observed. This implies that not all physicians interpret a non-diagnostic lung scan result in an identical way.

The calibration curve for the post-test probabilities shows good calibration for the categories 0-25 and 75-100%. (Figure 4) These categories consist mainly of patients with a normal or high probability V/Q scan result. Physicians were less well calibrated when assigning post-test probabilities to patients with a non-diagnostic scan, most of which ended up in the middle range of the probability scale. No significant differences were found in the mean pre-and post probabilities between residents and specialists.
Physicians’ interpretation of ventilation-perfusion lung scans

Figure 3 Scatter plots of pre-test and post-test probability estimates per test result for an overall of 271 patients. A) normal scan, B) non-diagnostic scan and C) high probability scan. The circles are from patients with confirmed pulmonary embolism and the cubes are from patients with no pulmonary embolism. (Non-interpretable scan, n=4, not shown)

Figure 4 Calibration curve comparing physician post-test probability estimate with the actual rate of pulmonary embolism (PE) for the 275 patients. For each point estimate, the 95% confidence interval is given. The dashed line indicates perfect calibration between observed rate of pulmonary embolism and the post-test probability estimate. Note that this figure consists of 9 groups instead of the 10 groups, as two deciles had to be combined in a single group.
Subjective likelihood ratio versus objective likelihood ratio

We compared the mean subjective likelihood ratios, calculated from pairs of post-test and pre-test probabilities, with the objective likelihood ratios, as obtained from a comparison of test results with those from the reference standard. (Table 2) The objective likelihood ratio for a normal scan was 0.046 (95% CI: 0.012 to 0.19). The mean of the subjective likelihood ratio was almost identical: 0.056 (CI: 0.038 to 0.082). For a high probability scan both likelihood ratios were far from unity: 20 (CI: 8.8 to 43) (objective) versus 26 (CI: 16 to 41) (subjective). The objective likelihood ratio for a non-diagnostic scan was 0.43 (CI: 0.31 to 0.62), significantly lower compared to the subjective likelihood ratio of 1.1 (CI: 0.77 to 1.5) for this result (p<0.01).

Table 2 The geometric mean of the subjective likelihood ratio versus the objective likelihood ratio per category of V/Q test result.

<table>
<thead>
<tr>
<th>Category</th>
<th>n</th>
<th>Subjective Likelihood Ratio</th>
<th>Objective Likelihood Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LR</td>
<td>CI</td>
</tr>
<tr>
<td>Normal</td>
<td>63</td>
<td>0.056</td>
<td>0.038 to 0.082</td>
</tr>
<tr>
<td>Non-diagnostic</td>
<td>119</td>
<td>1.1</td>
<td>0.77 to 1.5</td>
</tr>
<tr>
<td>High probability</td>
<td>89</td>
<td>26</td>
<td>16 to 41</td>
</tr>
<tr>
<td>Non-interpretable</td>
<td>4</td>
<td>0.75</td>
<td>0.14 to 4.0</td>
</tr>
</tbody>
</table>

The tendency in physicians to judge a non-diagnostic scan as literally being non-informative - although the objective LR suggests that it should lower the strength of pulmonary embolism suspicion - can also observed from figure 5, which illustrates the likelihood ratios for a non-diagnostic scan per center. Though not significant it seems that the subjective LR overall is higher than the objective LR. In center 1 the subjective likelihood ratio was significantly lower than unity (0.54 (CI: 0.33 to 0.90)) comparable to its corresponding objective likelihood ratios (0.42 (CI: 0.22 to 0.79)). This result shows that these physicians adjust for center-specific differences in definitions of the category non-diagnostic test result. The subjective likelihood ratios in the other centers ranged from 0.77 to 9.4, with objective likelihood ratios ranging from 0.16 to 0.72. There were no statistically significant differences between residents and specialists.
Discussion

This study evaluated physicians interpreting ventilation/perfusion lung scan results in the workup of patients with suspected pulmonary embolism. The pre-test probabilities of pulmonary embolism, based on information from history, physical and routine testing, varied widely but were well calibrated. Pre-test probabilities were appropriately updated if the results of the lung scan were either negative (normal scan) or clearly positive (high-probability scan). Substantially more variability was observed for non-diagnostic scans, sometimes leading to an increased and on other occasions to a decreased level of suspicion.

The pre-test calibration curve observed here is rather typical for physicians who in general underestimate the prevalence of pulmonary embolism in the low suspicion patients and overestimate the risk in patients with a high suspicion of pulmonary embolism. The calibration for the post-test probabilities was almost perfect for clearly negative or positive results of the lung scan.

A subjective likelihood ratio is an expression of the diagnostic impact of a lung scan result. For the normal and high probability scan the subjective likelihood ratios were similar to the objective likelihood ratios, indicating that the various physicians in this study updated the pre-test probabilities according to the diagnostic worth of these scan results. This was not the case for the non-diagnostic scan result. The mean subjective likelihood ratio was close to unity, indicating that on average physicians remain uncertain of the diagnosis of pulmonary embolisms. Yet in reality the non-diagnostic scan has diagnostic worth, as expressed by the objective likelihood ratio, which was significantly lower than one.

Variability in the probabilities attached to a non-diagnostic result could not be explained by differences between the centers. Here also the subjective
likelihood ratio tended to be higher than the objective likelihood ratio. Only for
center 1, the subjective likelihood was consistent with the objective likelihood
ratio. Both were significantly less than unity, indicating less pulmonary embolism
patients were assigned a non-diagnostic scan result than non-pulmonary embolism
patients. In this center physicians, knowing of earlier cases, did assign weight to a
non-diagnostic result.

Overall one can say that physicians intuitively updated probabilities in
accordance with the diagnostic worth of the lung scan. This finding underscores
Frybacks view that subjective probabilities are not just guesses, but the result of
incorporating all relevant information with the appropriate weight into an opinion.21
The physicians’ judgment refers to an intellectual process with incorporation of all
subjective and objective evidence with their appropriate weight using known
physiological and anatomical models relevant for the disease in question.
Nevertheless the conclusion that the physicians in this study properly adjust their
uncertainty can be challenged. Maybe they did so explicitly. Lung scanning is not
a new technique and it is possible that some physicians were using previously
published likelihood ratios. On the other hand, the quality of their judgment may
well be due to the large amount of short-term feedback they receive of the final
pulmonary embolism diagnosis. Such feedback mechanisms make physicians
perform better on familiar tests compared to novel tests.22

Probability estimates were available for only 275 patients. Although, in
principle, it is possible that physicians refrained from assigning probabilities to
more difficult patients, we do not think that this selection has caused a bias. On
the contrary almost all patients for which the reason for not assigning a posttest
probability was ‘not applicable’ had a normal scan. Possible in these cases a
pulmonary embolism as the cause for the complaints was refuted and another
diagnosis was considered giving the idea to the physicians that probabilities were
not necessary anymore. In one center 51% of the pre-test and 63% of the post-
test probability expressions were missing, indicative of the difficulties in obtaining
probability expressions. The prevalence of pulmonary embolism in this particular
group of patients with a normal scan as well as the overall prevalence are consequently
overestimated. Yet we think this did not affect our conclusions. Comparing the
subjective likelihood ratio to the objective likelihood ratio based on the 517 patients
who had a final diagnosis gave the same results.

A number of studies have previously reported objective likelihood ratios for
lung scan results. The objective likelihood ratios that can be derived from the
PIOPED study are comparable to the ones obtained in the study reported here.6 23

A small series of studies looked at the interpretation of lung scans by physicians.
One study found that a normal scan was misinterpreted: 31% of physicians remained uncertain about the diagnosis. In another study by the same authors, published in 1998, physicians were reported to interpret lung scans quite well if odds were used to express the likelihood of pulmonary embolism. The authors’ finding that physicians adjusted the odds of pulmonary embolism according the pre-test probability is in line with our findings. A recent study found a good agreement between 3 observers in assigning probabilities of pulmonary embolism after evaluating lung scans and clinical information.

Are subjective probability estimates to be preferred over decision rules within strategies for diagnosing pulmonary embolism? It is plausible to say that objective methods do not suffer from the variation that does exist between physicians. Nevertheless it is also known that in general decision rules do not travel well between populations and poor implementation may hamper the accuracy of formal decision rules. In addition various authors confirmed that (aided) subjective probability estimates perform as well as decision rules. With respect to strategies for diagnosing pulmonary embolism it has been shown that two formal decision rules expanded by Wells performed poorly. Another study showed that a decision rule performed more accurate when combined with clinical judgment. Although physicians’ were reasonable calibrated we did not study the discriminating abilities of their assigned pre-test probability of pulmonary embolism. Considering these arguments subjective probability assessment may be the right choice for use within diagnostic strategies for pulmonary embolism.

In conclusion we can say that pre-test probability estimates are properly adjusted to well-calibrated post-test probabilities by physicians responsible for the workup of patients suspected for pulmonary embolism. Physicians tend to interpret a non-diagnostic scan as a indifferent result despite the diagnostic worth it has in reality.

Acknowledgments
This study was supported from a grant from the Dutch Health Insurance Council

The following investigators, all in the Netherlands, were part of the ANTELOPE Study Group: Academic Medical Center, Amsterdam: B.J. Sanson, J.G. Lijmer, F. Turkstra, M.H. Prins, H.R. Büller. Leiden University Medical Center, Leiden: W. de Monyé, M.V. Huisman, P.M.T. Patrynama. Leyenburg Hospital, The Hague: M.J.L. van Strijen, G.J. Kieft. Slotervaart Hospital, Amsterdam: M.R. Mac Gillavry, D.P.M. Brandjes. University Hospital Free University, Amsterdam: P.J. Hagen, O.S. Hoekstra, R.A. Manoliu, P.E. Postmus. University Medical Center, Utrecht. I.J.C. Hartmann, P.F.G.M. van Waes, M.P.M. Stokkel, C.F. Melissant, J.D. Banga.
References
Physicians’ interpretation of ventilation-perfusion lung scans

Abstract

Background - MRI and CT have similar diagnostic accuracy in the diagnosis of patients with lumboradicular symptoms. It is not clear which test to use in daily clinical practice. The aim of this study was to assess the information content of both imaging modalities using the entropy statistic.

Methods - Eligible were patients (18-70 years of age) with radicular pain in whom operative treatment was considered. Findings from the patients' history and physical examination were presented to two physicians in a standardised format. Based on these descriptions both physicians independently assigned the probability of 7 different entities (3 levels, left or right and a category other) on separate visual analogue scales. In two separate sessions, they interpreted the images from CT and MRI. In each session, new probabilities were assigned again. By comparing the pre-test and post-test probabilities for each patient we calculated the information content for each test, in terms of entropy, using a random effects logistic regression model. In addition the physicians were asked after CT and after MRI how certain they were about operating patients and whether additional diagnostic tests were deemed necessary.

Results - The data on 59 patients were available for the analysis. MRI had a significantly higher information content than CT (0.61 bits and 0.73 bits versus 0.43 bits and 0.58 bits, p < 0.01). There was significantly less uncertainty after MRI in the decision whether to operate or not (p=0.003), but not one of the tests on average provided 100% certainty. In up to 49% of cases additional diagnostics were judged to be necessary.

Conclusion - MRI seems to contain significantly more information than CT in the diagnosis of patients with lumboradicular symptoms, leading to higher degrees of certainty in clinical decision making.
9 Computed Tomography versus Magnetic Resonance Imaging in the reduction of diagnostic uncertainty in patients with lumboradicular syndrome

Introduction
Despite several comparative studies of computed tomography (CT) and magnetic resonance imaging (MRI) in the diagnosis of patients with lumbosacral radicular pain it is still unclear which of the two imaging tests serves best for this indication. The reported sensitivities and specificities range from 62 to 90% for CT versus 60 to 100% for MRI, with specificities between 70 to 87% for CT versus 43 to 97% for MRI.1

It is well possible that both techniques suffice in terms of diagnostic accuracy. Research on the diagnostic and therapeutic impact of these techniques for patients suspected of lumbar disc herniation is scarce for MRI2 3 and even non-existent for CT. Diagnostic impact is the effect a diagnostic test result has on the degree of certainty a physician has regarding a particular diagnosis and future management.4 5 One of the measures to express this level of certainty quantitatively is entropy, a measure of information content derived from conventional information theory.6-8 The difference in before and after test entropy is an indication of the information a test holds and subsequently can be used to compare the diagnostic impact of tests. The aim of this study was to compare CT with MRI in the diagnosis of herniated nucleus pulposus, using the entropy statistic and to explore if any differences in entropy were related to differences in therapeutic decision making. We estimated entropy from probabilities assigned by the physician to every patient, expressing the likelihood of disc herniation at 6 particular locations: HNP on disc space level L3/L4 left and right, L4/L5 left and right, L5/S1 left and right. A seventh category was used for other diagnoses. A similar study has been conducted by Albeck and colleagues, using the pseudo regret function to measure the informativeness of tests.9

Patients and Methods

Patients
Data were obtained as part of a prospective study to evaluate the diagnostic process of patients with low-back pain, performed at a Dutch teaching hospital, between June 1999 and June 2000. Eligible were adult patients, between 18 and 70 years of age, with (incapacitating) radicular pain, in whom conservative treatment for at least 4 weeks had been unsuccessful. In these patients operative treatment was considered and diagnostic imaging is used in the decision making process. These criteria for inclusion reflect the consensus statement on diagnosis and treatment of lumbosacral root entrapment syndromes by the Dutch Society for Neurology.10 Excluded were patients who 1) had undergone previous back
surgery, 2) had already diagnosed lumbar disc herniation 3) could not give informed consent, 4) were pregnant, 5) had a contraindication for MRI investigation. The Institutional Review Board had approved the study protocol. Only informed and consenting patients were included in the study.

Diagnostic procedure
This study focused on the information content of the two imaging modalities, CT and MRI, conditional on the information already available through history and clinical examination.

Lumbar CT examinations were performed on a Philips Elscint TWIN CT scanner. Helical CT was made with 1.1 mm collimation and 0.5 mm increments from the level of L3 to the bottom of S1. The gantry angle was angled approximately through the disc space of L4-L5. Reformatted axial sections were made using both the soft tissue and bone windows of all scanned interspaces, precisely parallel to the interspaces to achieve an optimal delineation of the disk. In addition reformatted sagittal sections were made. The distance of the sections was 4.0 mm and the thickness of the sections was for soft tissue window 4.0 mm.

Lumbar MRI examinations were performed with a 1.5 T Signa Scanner (General Electric Company). The protocol included sagittal T1-weighted (TR 500, TE 14) spin-echo images and proton density T2-weighted (TR 3500, TE 120/20) fast spin-echo images with 4 mm slice thickness and 0.5 mm intersection gap; matrix 200x512; field of view 29 x 29 cm. In addition axial T1 (TR 520, TE 12)- and T2 (TR 4500, TE 120) -weighted fast spin-echo images were obtained from the level of L3 to the bottom of S1 with 4 mm slice thickness and 0.5 mm intersection gap; matrix 200x256; field of view 15 x 15 cm. Axial images were obtained without angulation. Finally, heavily T2- weighted (TR 5000, TE 252) fast spin-echo oblique MR myelography was performed with two slices of 20 mm thickness images; matrix 250 x 220; field of view 16 x 16 cm.

To standardise the clinical information the patients were presented as case descriptions to an experienced neurologist and a neurosurgeon. These experts then assigned seven probabilities on a visual analogue scale (VAS) for seven possible entities: hernia at disc space L3/L4 left or right, L4/L5 left or right, and L5/S1 left or right and ‘other’. In separate sessions, the physicians were presented with the same case descriptions, this time accompanied by the matching CT or MRI images with the accompanied report from the radiologist. The experts were asked to assign new probabilities on seven similar VAS scales. The two sessions were separated in time by at least one month. The two physicians were also asked to indicate how certain they were in making a decision on operative treatment and
whether other diagnostic tests were deemed necessary after interpreting the image tests.

Analysis

Entropy was used to measure the information content of the CT and MRI. This measure, developed within information theory, expresses information in bits. For a single condition that is either present or absent and a dichotomous test, an entropy change of 1 bit after the test result has been obtained would mean that the pre-test probability of 0.5 - reflecting maximal uncertainty - is updated to either 1 or 0, reflecting maximal certainty.7-11

To estimate the information content of CT and MRI the difference between the pre-test and post-test entropy was calculated for each patient. The pre-test entropy is estimated as

\[ H_{\text{pre}} = - \sum_{i=1}^{7} P(D_i) \log_2 P(D_i) \]

where \( D_i \) is one of the 7 entities and the post-test entropy as

\[ H_{\text{post}} = - \sum_{i=1}^{7} P(D_i | T) \log_2 P(D_i | T) . \]

The probabilities per entity were taken from the scales. As the seven categories are considered to be mutually exclusive, the corresponding probabilities were normalised to sum up to 100%.

We then calculated the mean change in entropy per observer over all patients both for MRI as well as for CT to express the information content of the two techniques.

To test the null hypothesis of similarity in information content for CT and MRI, we used a paired t-test for each observer separately. For the two observers jointly, we used a mixed analysis of variances model. Such a mixed model accounts for the information content of CT and MRI of the two observers being determined in the same patients.12 A similar analysis was performed on the uncertainty expressions in the decision to operate. Computations were done with the statistical software package SPSS version 11.

Results

Of the 64 consecutively included patients, 5 had to be excluded for the following reasons; \( n=3 \) claustrophobia and \( n=2 \) loss to follow up. The characteristics of the 59 patients available for the analysis can be found in table 1.
Of these patients 24 (41%) were operated on. The surgical findings were: 1 herniated nucleus pulposus (HNP) on the level lumbar disc 2 and 3 (L2-L3) left, 12 HNP’s L4-L5 (7 right, 5 left sided), 8 HNP’s L5-S1 (4 right, 4 left sided) and 3 patients with root canal stenosis. Of the patients that were assigned to conservative therapy 27 had a HNP on the imaging tests. In these patients the location of the HNP did not match the nature of the complaints or the severity of symptoms had decreased while waiting for surgery.

Figure 1 illustrates, for each patient, the entropy after CT or MRI (Y-axis) relative to the entropy prior to testing for that same patient (X-axis). Deviations from the diagonal reflect a change in entropy, standing for a change in uncertainty. Points above the diagonal reflect a decrease in certainty and points below the diagonal indicate an increase in certainty for the respective patients. Both tests seem to decrease uncertainty for most patients, with only a few points lying above the diagonal.

Table 2 shows the results of the formal analysis. In theory, uncertainty (entropy) is maximal when all possible entities have an equal probability. As we have 7 entities, this would imply seven times a probability of 0.14 (1/7), and the maximal entropy would be 2.8 bits. The pre-test entropy, based on the clinical information only, is close to 1 for both observers, indicating that most uncertainty is eradicated after history and physical examination.

<table>
<thead>
<tr>
<th>Females</th>
<th>39 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age, years (SD)</td>
<td>44 (12)</td>
</tr>
<tr>
<td>BMI (SD)</td>
<td>25 (3.6)</td>
</tr>
<tr>
<td>History of low back pain/ leg pain</td>
<td>74 %</td>
</tr>
<tr>
<td>Median duration symptoms, months (IQR)</td>
<td>3 (2–6.8)</td>
</tr>
<tr>
<td>Symptoms</td>
<td></td>
</tr>
<tr>
<td>Low back pain</td>
<td>88 %</td>
</tr>
<tr>
<td>Leg pain</td>
<td>98 %</td>
</tr>
<tr>
<td>Pain if ↑ abdominal pressure</td>
<td>78 %</td>
</tr>
<tr>
<td>Paresthesia</td>
<td>30 %</td>
</tr>
</tbody>
</table>
From Table 2, one can observe that MRI leads to a higher reduction in entropy than CT. The overall difference between the information content of CT and MRI incorporating the results of both observers was significant (p = 0.01).

Table 2 Results of the information content in bits.

<table>
<thead>
<tr>
<th></th>
<th>Observer S</th>
<th>Observer P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test entropy after H&amp;A</td>
<td>0.97 (0.044)</td>
<td>1.1 (0.037)</td>
</tr>
<tr>
<td>Post CT</td>
<td>0.54 (0.059)</td>
<td>0.53 (0.070)</td>
</tr>
<tr>
<td>Post MRI</td>
<td>0.35 (0.062)</td>
<td>0.38 (0.054)</td>
</tr>
<tr>
<td>Information CT</td>
<td>0.43 (0.072)</td>
<td>0.58 (0.079)</td>
</tr>
<tr>
<td>Information MRI</td>
<td>0.61 (0.076)</td>
<td>0.73 (0.063)</td>
</tr>
</tbody>
</table>

P = 0.008 for the difference between CT and MRI for observer S, and p = 0.066 for observer P (paired t-test)
Table 3 Results of the entropy change in the uncertainty to operate.

<table>
<thead>
<tr>
<th></th>
<th>Observer S Mean (se)</th>
<th>Observer P Mean (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial entropy after H&amp;A</td>
<td>0.61 (0.030)</td>
<td>0.39 (0.045)</td>
</tr>
<tr>
<td>Post CT</td>
<td>0.31 (0.043)</td>
<td>0.23 (0.040)</td>
</tr>
<tr>
<td>Post MRI</td>
<td>0.15 (0.032)</td>
<td>0.19 (0.035)</td>
</tr>
<tr>
<td>Information CT</td>
<td>0.30 (0.037)</td>
<td>0.15 (0.050)</td>
</tr>
<tr>
<td>Information MRI</td>
<td>0.46 (0.031)</td>
<td>0.20 (0.052)</td>
</tr>
</tbody>
</table>

P=0.01 for the difference between CT and MRI for observer S, and p=0.37 for observer P. (paired t-test)

Table 3 shows the results of the uncertainty in the decision to operate the patient or not. Physicians reported less uncertainty about operative treatment after MRI compared to CT (p=0.003).

We also asked the physicians if they needed additional diagnostic information for the decision to operate. To reach a decision on the need for surgery, observer S needed additional diagnostic information in 47% (28/59, all MRI’s) of patients after interpreting CT and in 3.4% (2/59, 1 MRI and 1 CT) of patients after interpreting MRI (p<0.001). These values were 49% (29/59, all MRI’s) and 12% (7/59, 4 MRI’s, 1 CT, 1 combination (MRI with intravenous contrast and CT) and 1 myelography) respectively for observer P (p<0.001).

Discussion
In this study the information content of CT and MRI was evaluated in the diagnosis of herniated nucleus pulposus in patients with lumbaradicular syndrome. We found MRI to yield more information than CT, leading to a higher reduction in uncertainty on the presence and the location of a herniated nucleus pulposus and a higher level of certainty in the decision to operate. Both observers needed additional diagnostic information in order to make this decision for almost half of the patients after CT. For MRI significant less additional diagnostic tests were needed. This underscores the finding that physicians are more confident after interpreting MRI relative to CT. This might be due to the higher resolution of MRI and as such a better contrast between the dural sac and herniation. The reason that for some patients after MRI a CT scan was needed might be due to uncertainty regarding possible bone abnormalities as a cause of the complaints.

Multiple studies have evaluated the difference in discriminating performance between MRI and CT. Recommendations based on the results from these studies
were equivocal in terms of which one is to be preferred over the other. Albeck and colleagues have previously compared the information content of the MRI, CT and myelography in diagnosing lumbar disc herniation.\textsuperscript{9} They concluded that CT produced more certainty relative to MRI, but that this difference was not statistically significant. Myelography yielded significantly lower information compared to CT and MRI.\textsuperscript{9} Myelography has become obsolete in our hospital, because of its invasiveness and the availability of MRI or CT. In contrast with the paper of Albeck, our study revealed that MRI held significantly more information than CT. This might be due to increased refinements in MRI technique relative to CT from the time Albeck and colleagues collected the data. We do not think our physicians were biased because of a pre-existing preference for MRI. The time interval between CT and MRI interpretation was long enough to prevent such memory bias in favour of MRI. Albeck and colleagues used the pseudo regret function. The interpretation of this measure is similar to the entropy measure. Both reach a maximum at maximal uncertainty, and are zero when uncertainty no longer exists. Since we incorporated the differential diagnosis consisting of 7 different entities, the entropy statistic was more convenient. In addition it was not needed to verify the true diagnosis in the patients.

We are aware of three studies that studied the impact of MRI on the management of patients with possible disc herniation.\textsuperscript{2,3,14} They all concluded that MRI had considerable impact on confidence in the diagnosis and patient management. One study found that despite this impact no effect was seen on patient outcome. Unfortunately none of these studies compared MRI with CT.\textsuperscript{14}

The design of the study did not allow us to evaluate the consequences of extra information content of MRI on patient outcome. It is possible that the difference in entropy is inconsequential, leading to similar health outcomes in the end. The results cannot be unconditionally generalized to other settings, as we relied on the readings of two observers only. A more stable conclusion could be drawn when for more observers a comparable result was found.

The reason this study was performed was to investigate which test to use within a guideline in the workup of patients suspected for herniated nucleus pulposus. Although MRI was superior in the information content we can't state with firm conviction that MRI should become the one and only test to do. Depending on resources like costs or availability, CT is an adequate alternative imaging test in diagnosing the possible cause of patients with lumboradicul syndrome.
Acknowledgements
We thank Kees Albrecht for his advise on the design of this study and we thank Frank Jan Hulsman for interpreting the images.

References
Methodological relevance of this thesis
This thesis applies and extends methodological tools that range from the field of diagnostic accuracy to the field of diagnostic impact studies. Diagnostic accuracy studies deal with how good a test distinguishes the sick from the well. Diagnostic impact refers to the effects of a test-result on the physicians’ confidence in a particular diagnosis.1 2

With regard to the diagnostic accuracy we encouraged the use and interpretation of the diagnostic odds ratio in chapter 1. By adding this test measure of discriminating performance to the set of available measures we feel a comprehensive toolkit of accuracy measures is available to address the many study questions possible when evaluating the discriminating performance of diagnostic tests.

The most frequently used pair of outcome measures is the combination of sensitivity and specificity. Yet estimating these measures is not clear cut in all situations. In chapter 2 a method is proposed to assess a per patient sensitivity and specificity for tests that aim to detect multiple lesions of which the localization is important. We applied this method to a study designed to evaluate different levels of radiation dose in the detection of polyps by computed tomography colonography (CTC) in patients at risk for colorectal carcinoma (chapter 4). A comparison between different levels of radiation dose based on the per patient sensitivity and specificity could otherwise not have been made. One of the limitations of this method is the somewhat different interpretation of sensitivity and specificity. Furthermore, no diagnostic odds ratio can be calculated from the given sensitivity and specificity. Especially for this application - a comparison between different dose levels - the odds ratios would have been handy.

Synthesizing results from different (diagnostic) studies in systematic reviews can be challenging since many issues have to be considered. One of these is the clinical and methodological heterogeneity between studies. The technique used to synthesize results depends upon the type and the amount of heterogeneity present between studies.3 4 When the aim of the meta-analysis is to pool results of tests so to compare these different tests, difficulties arise when different methods have been applied for pooling each test. In chapter 2 a bivariate model is suggested that can address most sources of heterogeneity. An advantage of this approach to meta-analysis is that it gives summary estimates of sensitivity and specificity, the statistics most often used to characterize diagnostic accuracy.

In chapter 5, the bivariate model is applied to pool the results of six different diagnostic tests with the aim see which one is best. When meta-analytic methods will be used for comparing tests it is possible that the results from specific tests are retrieved from comparative studies. In that case a dependency will exist between
the results for tests that have been applied in the same population. In our comparison we did not take this into account. Research is needed to evaluate the magnitude of this dependency and to decide when it is needed to incorporate an extra variable in the model to deal with this dependency.

As mentioned in the general introduction accuracy measures fail to address the added value of a test or the impact a test has on the likelihood of disease. This theme brings us into the field of diagnostic impact or diagnostic thinking efficacy studies.1 2 In chapter 6 to 9 subjective probability estimates were used to address such questions. These probabilities express the degree of belief a physician has in a particular diagnosis. They were assigned by physicians on a double logarithmic visual analogue scale.

In chapter 6, subjective probability estimates were used as prior probabilities to study whether probabilities were consistently updated after the test results became available. We were able to use these pre-test probability estimates to identify subsets of patients where the test has added value.

In chapter 7 the probability estimates were used to compare clinical judgment with objective decision tools. Chapter 8 contains a report of a study on how well physicians interpret the ventilation-perfusion lung scan. This was done by calculating the subjective likelihood ratio (LR) from the pre and post-test probability estimates assigned by physicians. Subsequently these subjective LR were compared with the objective LR, obtained form a comparison of the test results with those of the reference standard.

In chapter 9, a transformation of the subjective probabilities was used to specify the amount of information in a test result. This value - entropy - is a measure derived from information theory. It was used to estimate the information content of two tests so to compare them with one another. Entropy allows us to incorporate the information the test provides about the target disease as well as on alternative diagnoses. In addition it is not necessary to define a test result as positive or negative, as was the case in the diagnoses of herniated nucleus pulposus, where there is no unequivocal way to define a negative or positive test result. Whether a test result is labeled as positive will depend on whether the finding on the imaging test matches the pattern of signs and symptoms a patients has, as well as on the result of the reference standard. In our study, physicians were only asked to indicate whether their uncertainty had changed after the results of the imaging test had become available. The entropy measure is seldomly used in diagnostic studies. This might be due to difficulties in interpreting this measure. Further research is needed to explore the usefulness of such measures.
Clinical relevance of this thesis
Although this thesis focuses on methodology, we feel the results of the studies reported here also have clinical value.

In chapter 4, it is shown that low dose scanners have equal diagnostic accuracy compared to the currently used scanners. At low doses the risk of radiation induced cancer is lower. With these results hopefully manufactures will be convinced to develop low dose scanners.

In the study reported in chapter 5 we found that most urine-based tumor markers in the detection of bladder cancer lack diagnostic accuracy. In this light cystoscopy should stay the instrument of choice to rule out bladder cancer. In chapter 6 we reported on a study on the added value of the dobutamine stress echocardiography test in patients with chest complaints but with ruled-out acute cardiac disease. This study showed that, although the test has predictive value, the added overall value is low. It is possible the test will perform adequately only in specific subgroups. In chapter 7 clinical judgment was compared with the Ottawa ankle relay clinical decision rules. We illustrated that physicians do as well as objective decision rules in ruling out fractures in patients with sprained ankles, and that they do better in saving unnecessary radiographs. Chapter 8 showed that physicians do well in interpreting ventilation perfusion scans, except for the non-diagnostic test results, where much variation exists in physicians’ interpretation. In chapter 9 we found that magnetic resonance imaging holds more information than computed tomography in diagnosing herniated nucleus pulposus, although both tests required additional diagnostic information to decide whether to operate upon a patient.

Diagnostic Framework
Ever since Yerushalmy proposed to use the sensitivity and specificity as outcome measures in the evaluation of a tests discriminating performance - in 1947 - many efforts have been put into the development of a framework to evaluate new diagnostic technologies. This movement was further stimulated by the notion that medical practice, including the implementation of new technology, needs to be Evidence Based. Such a framework should act analogous to the four phase model used to evaluate new therapies.

So far we have described the results of this thesis in the framework as proposed by Fryback et al. This framework is based on six hierarchical levels a test needs to go through before it should be used in clinical practice. The levels are: technical efficacy (technical properties of the test), diagnostic accuracy efficacy, diagnostic thinking efficacy, therapeutic efficacy (does the test in potential change the
management plan), patient outcome efficacy (does a test increase quality of life) and societal efficacy (cost/benefit evaluations).

None of the proposed frameworks has succeeded in establishing a central role in the process of developing and implementing new diagnostic tests. This is due to many factors, involving both the designs and outcomes used in test research, as well as the nature of diagnostic tests themselves. Many different designs and outcomes are possible depending on the different research questions that can be formulated. Examples are: what is the accuracy of test X, which test, from a series of tests, is best in terms of sensitivity and specificity, does test X have the same accuracy as test Y, what is the best cut-off value of continuous test X, how does test X perform in population K, what does this test add relative to the diagnostic information already available. Detecting a simple hierarchy in these questions is a strenuous task.

Diagnosis in itself is not that straightforward as therapy, where one aims at influencing the course of disease to reduce the patients complaints. The function of diagnostic testing can be described as to reduce the uncertainty of the physician in particular diagnosis. This is often a stepwise process of multiple testing involving many clinical factors. In addition diagnostic tests develop in an enormous rate. Before a test is evaluated well, an even newer version has come up.

A workable framework should be extremely flexible. A suggestion is a framework iterative in character, included the following phases:
- Step 1: Development of a test.
- Step 2: Understanding the single test in clinical practice.
- Step 3: Test within a diagnostic strategy.
- Step 4: Implementation, evaluation and evolution.

The research described in this thesis would fit in step 2, except perhaps for chapter 8. This chapter focused more on the evaluation of a specific test, in this case the ventilation-perfusion scan. Based on this research, we found that substantial variation exist in the interpretation of a non-diagnostic test result. These findings could have consequences for the test strategy of diagnosing pulmonary embolism.

We feel that it is important that a workable framework is developed in the near future. This should go together with developing an overview of typology of possible research questions with their designs. Current (lack of) resources does not permit the implementation of poorly evaluated diagnostic tests, which in the end might waist money and even worse might harm patients. Many hospitals already develop guidelines when to implement new diagnostic technology. It would be advantageous if a framework was available where a test should pass through before
it is accepted on the market. Researchers in the field should take the initiative to develop a useful framework, but governmental bodies (and health insurance companies) should provide the means to clear the way for medical researchers.

Clinical Judgment versus Decision Rules
In chapter 7 we found that physicians performed as well as two clinical prediction rules in ruling out fractures in patients with acute ankle injury. Later on in chapter 8 we found that physicians did well in overall interpreting the ventilation-perfusion scans of patients suspected for pulmonary embolism. These findings illustrate that in specific situations physicians are able to make excellent clinical decisions, outperforming formal statistic tools.

On the other hand, in clinical practice several initiatives have been deployed to construct decision aids. These objective tools are thought to lead to less variation in clinical practice relative to clinical judgment in making decisions.\cite{9} In the end formal decision aids will yield less medical errors. So the underlying rationale of the development of decisions aids is that clinical judgment does not suffice in making proper medical decisions. One other reason for an increase in decision tools is that the broader availability and knowledge of statistical regression and neural network techniques stimulated the production of such rules.

Yet it is also shown that such decision aids lack generalizability to other but related populations compared to the populations where they were derived from.\cite{10,11} Another disadvantage is that implementation of decision rules so far is expensive and often unsuccessful.\cite{12,13} In the future, more research is needed to answer the question in what situations decision rules might actually be valuable. This should prevent a waste of time, energy and money in the development, implementation and evaluation of prediction rules where actual physicians’ decision making skills are more than adequate.

In addition research should be undertaken to unravel the processes of doctoring itself. To do so, physicians could embrace results from cognitive psychology, the science that studies clinical judgment. A proper awareness of the products of this science might accomplish that physicians anticipate better, avoid common pitfalls, understand repeated mistakes, so they will be more cautious in daily clinical practice.\cite{14}

In the end an alternate use of decision rules with clinical judgment on evidence based ground will result in patient benefit as well as in a more intelligent use of resources.
References
In het huidige tijdperk van evidence based medicine, waarbij beslissingen worden genomen op basis van het beste beschikbare ‘bewijsmateriaal’, is het gewenst dat de effectiviteit van bestaande en nieuwe diagnostische tests wetenschappelijk worden onderzocht. Het meest bekende evaluatiemodel daartoe omvat zes sequentiële fases. In fase 1 wordt de test technisch ontwikkeld; in fase 2 vindt onderzoek plaats naar het onderscheidend vermogen van de test; fase 3 evalueert de impact die de test heeft op de klinische praktijk; de onderzoekingen in fase 4 en 5 richten zich respectievelijk op het effect van tests op therapie en patiënt uitkomsten; in fase 6, tot slot, worden de kosten en baten van de test vanuit een maatschappelijk perspectief beschouwd.

De hoofdstukken 1 tot en met 5 van het onderhavige proefschrift hebben betrekking op de tweede fase van het model. In deze fase wordt het onderscheidend vermogen van een test onderzocht. Het gaat dan om de vraag, hoe goed een test een patiënt met de betreffende ziekte kan onderscheiden van een patiënt die de verwachte ziekte niet heeft. De klassieke manier om dit na te gaan is de te onderzoeken test (index test) toe te passen op een groep patiënten die verdacht werden van de ziekte. Daarna worden deze patiënten onderworpen aan een zogenaamde referentie test. Dit is de test die op dat moment de optimale methode is om uitsluitend te geven welke patiënten daadwerkelijk ziek of niet-ziek zijn. De resultaten van beide tests worden vervolgens samengevat in een 2x2 tabel. (zie figuur)

Met behulp van de celfrequenties in de tabel kan men vervolgens langs een rekenkundige weg de uitkomstmaten bepalen voor het onderscheidend vermogen van de test. De sensitiviteit (TP percentage) en specificiteit (TN percentage) zijn de meest gebruikte uitkomstmaten. De sensitiviteit geeft de kans weer op een positieve test uitslag bij zieke patiënten. De specificiteit drukt de kans uit op een negatieve test uitslag bij de niet-zieke patiënten.

Een andere uitkomstmaat is de zogenaamde diagnostische odds ratio (DOR). In hoofdstuk 1 wordt beschreven voor welke diagnostische toepassingen de DOR geschikt is en hoe de DOR geïnterpreteerd kan worden. In tegenstelling tot de sensitiviteit en specificiteit, die alleen paarsgewijs geïnterpreteerd kunnen worden, is de DOR een enkelvoudige maat. Dit vereenvoudigt een snelle vergelijking van verschillende diagnostische tests. Een nadeel van de DOR is dat de afzonderlijke fout percentages onder de zieken en niet-zieken niet beoordeeld kunnen worden.

In hoofdstuk 2 wordt een methode beschreven die de sensitiviteit en specificiteit kan schatten bij tests die meerdere bevindingen per patiënt kunnen detecteren. Een voorbeeld van zo’n test is CT colonography, welke poliepen kan detecteren in de dikke darm. Patiënten kunnen in hun darm meerdere poliepen
Samenvatting

hebben. Het is de vraag welke bevindingen gevonden met behulp van CT colonography echte poliepen zijn. Het kunnen namelijk ook artefacten zijn. Als een patiënt bijvoorbeeld een echte poliep en een artefact blijkt te hebben, is het lastig deze patiënt te rangschikken in de 2x2 tabel. Met behulp van het beschreven wiskundig model kan toch een sensitiviteit en specificiteit berekent worden voor de CT colonography.

Hoofdstuk 3 gaat in op een meta-analytische methode die het mogelijk maakt om verschillende diagnostische onderzoeken met betrekking tot een bepaalde test samen te vatten en weer te geven in termen van sensitiviteit en specificiteit. De methode is gebaseerd op een bivariaat normale verdeling. Het model biedt de mogelijkheid om op eenvoudig wijze ‘random effects’ parameters op te nemen zodat op die manier statistisch gecorrigeerd kan worden voor bronnen van variatie. Een bijkomend voordeel van deze methode is dat zij met behulp van standaard statistische software (SAS) kan worden uitgevoerd.

In hoofdstuk 4 wordt een onderzoek beschreven naar het onderscheidend vermogen van verschillende gesimuleerde doses van radioactieve straling die nodig zijn bij het maken van CT colonography (CTC) beelden. De vraag was of lagere doses net zoveel poliepen kunnen detecteren als de normale dosis. Met behulp van de methode uit hoofdstuk 2 werd per dosis de sensitiviteit en specificiteit voor CTC geschat. Met uitzondering van de allerlaagste dosis, blijkt het onderscheidend vermogen gelijk te zijn.

Hoofdstuk 5 richt zich op een meta-analyse van tests die tumor merkstoffen kunnen detecteren uit de urine van patiënten met blaaskanker. Deze meta-analyse maakt gebruik van de methode zoals beschreven in hoofdstuk 3. Hoewel er veel dergelijke tests in omloop zijn, kan op basis van de wetenschappelijke literatuur geen ‘evidence based’ besluit worden genomen welke test het meest geschikt is. Uit onze meta-analyse kwam naar voren dat cytologie de beste specificiteit heeft en telomerase relatif de beste sensitiviteit. Helaas zijn beide testen onvoldoende sensitief om blaaskanker uit te sluiten. Geen van de onderzochte testen is dan ook geschikt voor de klinische praktijk.

Vanaf hoofdstuk 6 schenken de bestudeerde onderwerpen aandacht aan de derde fase van het evaluatiemodel, namelijk de fase van de diagnostische impact. Hier vindt een evaluatie plaats naar de toegevoegde waarde van de diagnostische test. Een test met een goed onderscheidend vermogen hoeft namelijk nog geen grote impact te hebben op de klinische praktijk. Het doel van diagnostische tests is immers het verminderen van de onzekerheid van de dokter (en van de patiënt) over de aanwezigheid of ernst van een bepaalde aandoening. In het geval de arts al
vrijwel 100% zeker is over zijn of haar diagnose dan zal een test, hoe goed deze moge zijn, nauwelijks meer een bijdrage leveren aan de (differentiaal) diagnose.

Hoofdstuk 6 beschrijft een onderzoek naar het gebruik van de dobutamine stress echocardiografie (DSE) bij patiënten die via de eerste harthulp opgenomen waren, maar geen aantoonbare hartproblemen hadden. Het doel van de DSE is te voorspellen of deze patiënten een verhoogde kans hebben op een hartaanval binnen 6 maanden na ontslag. DSE bleek over het algemeen weinig bij te dragen aan het vermoeden van de arts over de prognose van de patiënt. Dit vermoeden werd uitgedrukt in een kansschatting op basis van diagnostische informatie die tijdens de opname was verkregen. Uit een aanvullende analyse kwam naar voren dat de test wel van waarde was bij een door de arts geëvalueerde subgroep van patiënten met een matig risico.

Ook in hoofdstuk 7 werd aan artsen gevraagd kansen te schatten. Het doel van dit onderzoek was het klinisch oordeel van artsen te vergelijken met twee objectieve beslisregels. In veel ziekenhuizen wordt standaard een röntgenfoto aangevraagd bij patiënten met een verstuwde enkel. Omdat slechts een gering aantal van deze patiënten (<10%) een fractuur hebben, zijn indertijd beslisregels ontworpen om de aanvraag van röntgenfoto’s terug te dringen. Aan Eerste Hulp artsen werd gevraagd hoe groot de kans was op een gebroken enkel bij mensen met een verstuwde enkel. Tevens moesten zij aangeven of zij het nodig vonden een röntgenfoto te laten maken. Uit het onderzoek kwam naar voren dat artsen even goed als de beslisregels konden bepalen of een röntgenfoto nodig was om een gebroken enkel uit te sluiten. Ook bleken artsen minder röntgenfoto’s aan te vragen in vergelijking tot de objectieve beslisregels.

In hoofdstuk 8 wordt een onderzoek beschreven naar wijze waarop artsen ventilatie-perfusie scans beoordelen om de diagnose longembolie te stellen. Aan artsen werd gevraagd om voor en na de long scan de kans op een longembolie te bepalen. Vervolgens werd met behulp van de regel van Bayes voor elk van de drie mogelijke testuitslagen een subjectieve likelihood ratio berekend. De likelihood ratio is een maat die aangeeft hoe vaak een bepaalde test uitslag voorkomt bij zieke patiënten ten opzichte van niet-zieke patiënten. Deze subjectieve likelihood ratio werd vervolgens vergeleken met de objectieve likelihood ratio, die werd berekend aan de hand van een vergelijking met de referentie test. Het bleek dat artsen een normale test - en een hoge kans uitslag naar waarde interpreteren. Daarentegen was er onder de artsen veel onderlinge variatie in de beoordeling van een non-diagnostische test uitslag. Juist deze test uitslag komt het meest voor bij patiënten die verdacht worden van een long embolie.

In hoofdstuk 9 wordt een vergelijking gemaakt tussen de CT en MRI scan
voor de diagnose van hernia nucleus pulposus bij mensen met lage rugpijn met uitstraling naar de benen. Omdat uit eerder onderzoek was gebleken dat er waarschijnlijk geen verschil is in onderscheidend vermogen tussen CT en MRI in deze patiëntengroep werd een maat gekozen die de informatiewaarde van een test kan uittrekken. De informatiewaarde van een test kan opgevat worden als het verschil in onzekerheid van een arts voor en na het bekend worden van het testresultaat. In deze studie werd gekozen voor de entropy, een maat afkomstig uit de informatietheorie. Het resultaat van onze evaluatie was dat de MRI meer informatie oplevert dan de CT scan, ook ten aanzien van de beslissing om de patiënt al dan niet te opereren. Echter voor beide tests gold wel dat in veel gevallen nog aanvullend onderzoek noodzakelijk was.

Zoals eerder opgemerkt, beschrijft dit proefschrift de ontwikkeling en toepassing van methodologie voor het evalueren van diagnostische tests in de tweede en derde fase van het evaluatieonderzoek. Het doel van het sequentiële fasemodel (en varianten daarop) is de evaluatie van diagnostische tests te structureren. Een dergelijke stapsgewijze structurering is analoog aan de verschillende fasen die een nieuw medicijn moet doorlopen alvorens zij kan worden geregistreerd en geïntroduceerd in de reguliere gezondheidszorg. Het karakter van diagnostische tests, hun doel en de diagnostische teststrategieën complicerken echter een sequentieel verlopend evaluatieproces. Dit is waarschijnlijk de reden waarom het succesvol doorlopen van de diverse fasen nog geen voorwaarde is voor succesvolle implementatie in de praktijk. Een flexibel systeem met cyclische kenmerken zou wellicht meer passen bij het specifieke karakter van diagnostische test evaluaties. Onderzoek naar passende evaluatiemodellen is dan ook nodig zodat tests - meer dan nu - doelmatig worden toegepast en om te voorkomen dat patiënten schade ondervinden van niet goed geëvalueerde diagnostische procedures.

Uit de hoofdstukken 6, 7 en 8 komt naar voren, dat artsen in bepaalde klinische situaties goed medische beslissingen kunnen nemen, en soms zelfs tot betere besluiten in staat zijn dan objectieve beslisregels. Beslisregels worden in toenemende mate ontwikkeld om variatie in klinisch handelen te verminderen. Het blijkt echter dat beslisregels meestal minder goed presteren in patiëntengroepen waarin ze niet zijn ontwikkeld. Onderzoek is nodig naar het identificeren van klinische situaties waar beslisregels artsen van dienst kunnen zijn. Daarnaast dient meer verklarend onderzoek te komen naar de wijze waarop artsen beslissingen nemen. Zo krijgen artsen meer inzicht in hun eigen besliskundig proces, kunnen ze beter anticiperen op valkuilen in hun denken, en kunnen - als direct gevolg daarvan - medische fouten worden verminderd.
List of Publications


List of Publications


Glas AS, Bossuyt PMM. ‘De klinische betekenis van het aantonen van antikernantistoffen’. (brief) NTvG 2000; 144(22): 1087

Glas AS, Lijmer JG. Minder röntgenfoto’s en toch goede klinische zorg door geprotocolleerde fysische diagnostiek bij enkelletzels’. (brief) NTvG 2000; 144(30): 1458-1459

List of Co-authors

Rhada Bholasingh MD, Department of Cardiology, Academic Medical Center, Amsterdam, The Netherlands
Kjell Bogaard MD, Department of Orthopedic Surgery, Academic Medical Center, Amsterdam, The Netherlands
Gouke J. Bonsel MD PhD, Department of Public Health, Academic Medical Center, Amsterdam, The Netherlands
Patrick M.M. Bossuyt PhD, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, The Netherlands
Harry R. Büller MD PhD, Department of Vascular Medicine, Academic Medical Center, Amsterdam, The Netherlands
Rudolf M.J.M. Butzelaar MD PhD, Department of Surgery, Sint Lucas Andreas Ziekenhuis, Amsterdam
Marije Deutekom MSc, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, The Netherlands
Jasper Florie MD, Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands
Constantine A. Gatsonis PhD, Center for Statistical Sciences, Brown University, Providence, United States
Gerard J. den Heeten MD PhD, Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands
Johannes N. Keeman MD PhD, Department of Surgery, Sint Lucas Andreas Ziekenhuis, Amsterdam
Nina M. Klemetsø MD, Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands
Karl H. Kurth MD PhD, Department of Urology, Academic Medical Center, Amsterdam, The Netherlands
Johan S. Laméris MD PhD, Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands
Jeroen G. Lijmer MD PhD, Department of Psychiatry, Utrecht Medical Center, Utrecht, The Netherlands
Rogier G. van Gelder MD, Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands
Melvin R. MacGillavry MD PhD, Department of Internal Medicine, Slotervaart Ziekenhuis, Amsterdam, The Netherlands
Charles B.L.M. Mojoie MD PhD, Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands
List of Co-authors

C. Yung Nio MD, Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands
Wilco C. Peul MD, Department of Neurosurgery, Leiden University Medical Center, Leiden, The Netherlands
Bas A.C.M. Pijnenburg MD, Department of Orthopedic Surgery, Academic Medical Center, Amsterdam, The Netherlands
Martin H. Prins MD PhD, Department of Epidemiology, University Maastricht, The Netherlands
Johannes B. Reitsma MD PhD, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, The Netherlands
Jeroen C. van Rijn MD, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, The Netherlands
Daphne Roos MD, Department of Urology, Academic Medical Center, Amsterdam, The Netherlands
Marnix A.J. de Roos MD, Department of Surgery, Martini Ziekenhuis, Groningen, The Netherlands
Anne W.S. Rutjes MSc, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, The Netherlands
Michiel P. Schutter, Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands
Iwo W.O. Serlie MSc, Department of Applied Physics, Technical University, Delft, The Netherlands
Rob J.P.M. Scholten MD PhD, The Dutch Cochrane Center, Amsterdam, The Netherlands
Jan Stam MD PhD, Department of Neurology, Academic Medical Center, Amsterdam, The Netherlands
Jaap Stoker MD PhD, Department of Radiology, Academic Medical Center, Amsterdam, The Netherlands
Henk W. Venema PhD, Department of Medical Physics and Radiology, Academic Medical Center, Amsterdam, The Netherlands
Frans M. Vos PhD, Department of Radiology, Academic Medical Center, Amsterdam and Department of Applied Physics, Technical University, Delft, The Netherlands
Robbert J. de Winter MD PhD, Department of Cardiology, Academic Medical Center, Amsterdam, The Netherlands
Petrus M. van der Zee MD, Department of Cardiology, Academic Medical Center, Amsterdam, The Netherlands
Aeilko H. Zwinderman PhD, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, The Netherlands
Dankwoord

Iedereen die de afgelopen 4 jaar direct of indirect meegewerkt en gedacht hebben over de inhoud van dit proefschrift. BEDANKT!!

In het bijzonder:

Patrick, het was een ongekende luxe om jou als directe begeleider te hebben. Wat heb ik veel van je geleerd en wat is er veel dat ik nog van je kan leren. Ik hoop in de toekomst nog vaak met je samen te werken. Mijn allergrootste dank!!

Koos, de vordering van het onderzoek raakte na jou komst in een stroomversnelling. Je bent de katalysator achter het werk, ondanks de tijd die ik kwijt was door de door jou ingevoerde koffiepauze, die zo nu en dan flink uitliep. Met jou durf ik elk statistisch model aan. Heel veel dank!!

Hans, pas tijdens mijn eindspurt zijn we gaan samenwerken. Voor mij ben jij het voorbeeld van hoe arts-epidemiologen een brug moeten slaan tussen kliniek en biostatistiek. Bedankt voor het inzicht dat ik daardoor heb gekregen in de methodologie van diagnostische meta-analyses. Jij weet als geen ander hoe artikelen ontrafeld moeten worden.

Diagnostiek collega’s van kamer J2-205: Jeroen, bedankt dat je me de basis beginselen van het diagnostiek onderzoek hebt geleerd en me op weg hebt geholpen met het onderzoek. Melvin, toen je kwam werd ik overdonderd door je enthousiasme voor de wetenschap, elke dag minstens één nieuwe hypothese dat was zeker wennen. Al gauw werd ik aangestoken door je drive. Ik vond het heerlijk met je samen te werken en te discussiëren over methodologische problemen. Anne, jou werkschema, waardoor je vaak ‘s avonds nog werkt was een goede motivatie voor mij om ook ‘s avonds door te gaan werken. Zo heb je me door de laatste loodjes heen gehaald. Mijn dank daarvoor!! Heb je je contact met Dirk al uitgediept? Jeroen van Rijn, altijd in voor een pittige discussie, overigens niet altijd gewenst in tijden van werkstress. Je enthousiasme en heldere denken maken me nu al nieuwsgierig naar je toekomstige proefschrift, veel succes!

Bas, ik vond het erg leuk met je samen te werken. Het wordt weer hoog tijd voor een dubbel.

Collega’s van de radiologie: Shandra, Rogier, Frans, en Nina. Door onze samenwerking heb ik geleerd dat methodologische theorie in de praktijk vaak heel
 anders gaat en dat specifieke klinische vraagstellingen vaak een eigen benadering nodig hebben. Een loopje naar de radiologie was altijd een aangename afleiding. Bedankt voor de prettige samenwerking en heel veel succes met jullie onderzoek.

Marc van der Zee en Rob de Winter, bedankt dat ik jullie gegevens mocht gebruiken. Ik heb veel geleerd van de nuttige discussies en jullie uitleg betreffende de diagnostiek van pijn op de borst.

Alle mede auteurs van artikelen, in het bijzonder: Prof. Kurth, Gouke Bonsel. Martin Prins, Jaap Stoker, Daphne Roos, en Prof. Bøller, bedankt voor jullie hulp en kritische commentaren.

Medewerkers van de KEB, bedankt! Wat heb ik een ontzettend leuke tijd gehad op de KEB. Ik mis jullie en hoop nog vaak terug te komen. Lieve Gre, bedankt voor alle zorg en aandacht die ik kreeg met name tijdens die momenten dat ik even wat afleiding nodig had. Noor, ik mis je lach en felle blik. Hans, Carlo, Marije en Rebecca laten we nog vaak biertjes gaan drinken in de pijp. Joke, Gerrit-Jan, Tessa, Barbara, Michel, Marcel, Astrid, Brent, Corianne, Jeanette, Rebecca H, Robert, Rien, Peter, Miranda, Rob, Hanni, en Martin het was erg gezellig met jullie werken. Henk en Jos, jullie zijn onmisbaar. Bedankt voor het verhelpen van alle computerongemakken.

Verder wil ik bij deze nog bedanken Prof. dr. Gerjan Navis en dr. John Plukker die in Groningen aan de basis hebben gestaan van mijn wetenschappelijke carrière. Zonder jullie was ik nooit op het idee gekomen promotieonderzoek te gaan doen.

Mijn nieuwe collega’s in de Heel, bedankt dat jullie mij na 4 jaar kantoorbaan goed hebben opgevangen in de kliniek.

Lieve Knickerbocker vriendinnen bij Hercules: Gea, Sylvia, Ellen, Marije, Marianne, Brenda, Jose, Margriet, Janke, Marielle, Alie en Judith, mijn grote dank voor alle gezelligheid rond gewonnen en verloren wedstrijden en niet te vergeten de weekendjes.

Wartburgia dames 4+1: Andrea, Neála, Lara, Ellen, Sylvia, Sandra, Sannie, Gabi, Theresa, Barbara, Stien, Hedi, Carolien, Ellen, Griselda, Kitty en Leo en natuurlijk ook Joyce, Natascha, Hulda en Stef; een uniek team. Met jullie wordt
elke zaterdag een feest, ondanks Utrechtse tijden en de vele tegendoelpunten.

Hillie, Yvonne en Coralie, de afstand Gronigen-Amsterdam is ook figuurlijk vaak moeilijk te overbruggen. Bedankt voor de leuke momenten van ontspanning, ik hoop in de toekomst vaker contact met jullie te hebben.

Ellen Brand, het is heerlijk met jou wilde plannen te smeden, ik weet zeker dat we er een paar tot uitvoer zullen brengen.

Grace, partner in crime, now we both (almost) have our PhD, I think it’s time again to smoke some Cuban cigars and have some advokaat.

Lieve Heit en Mem, Ineke, Anita, Ruurd, Jouke, Harry en Marlies, bedankt voor jullie onvoorwaardelijke steun en vaste vertrouwen in wat ik doe, al heb ik nooit echt goed kunnen uitleggen waar ik precies mee bezig ben.

Lieve Siebe en Afke, als kind liep ik bij jullie de deur plat en heb onder andere geleerd dat de wereld groter is dan het dorp waarin wij woonden. Heel erg bedankt voor alle wijze lessen.

Jose, al heb je dit proefschrift vaak verwenst, bedankt voor al je liefde en geduld.

Sylvia en Ellen, ik ben erg vereerd jullie als mijn paranimfen te hebben. Met jullie durf ik de aanval aan!

Afina
Afina Glas was born on December 2nd 1971 in the town of Dokkum, the Netherlands. After graduating from the MAVO, HAVO and VWO in 1991, she studied for one year at Dordt College, Sioux Center, United States, where she obtained an Associate of Arts degree. In 1992 she started medical school, at the University of Groningen and achieved her degree in December 1998. Soon after that she started research in the field of methodology of diagnostic test evaluations at the Department of Clinical Epidemiology and Biostatistics of the Academic Medical Center in Amsterdam. This thesis is based on this work. In June 2001 she got her Masters of Science degree in Clinical Epidemiology at the Netherlands Institute for Health Sciences, Erasmus University, Rotterdam. In February 2003 she started with her specialization in urology at the University of Amsterdam (head: Prof. Dr. J.J.M.C.H. de la Rosette). Currently she works at ‘De Heel, Zaans Medisch Centrum’, in Zaandam as a resident in general surgery as part of the urology training (supervisor: Dr. A.E. Engel).