Evaluation of diagnostic tests: from accuracy to outcome
Lijmer, J.G.

Citation for published version (APA):
Non-invasive diagnostic tests for pulmonary embolism:
Less useful when needed the most.

Jeroen G. Lijmer, Melvin R. Mac Gillavry, Marco van Strijen,
Martin H. Prins, Dees P.M. Brandjes, Harry R. Büller on behalf of the
ANTELOPE Study group

Submitted
Abstract

Purpose The diagnostic work-up for patients suspected of pulmonary embolism (PE) consists of perfusion-ventilation lung scan (VQ scan) followed by a pulmonary angiography if the results remain inconclusive. Non-invasive diagnostic tests, such as D-dimer measurement, compression ultrasonography (CUS) and Spiral CT are advocated to reduce the number of angiograms in patients with a non-diagnostic VQ scan based on decision analyses. To assess the validity of the independence assumption in these decision analyses, we examined the diagnostic accuracy of non-invasive tests in patients with different VQ scan results.

Methods Data were obtained in a multicentre study in patients with suspected PE (n=627). PE was excluded by a normal pulmonary angiogram or a normal VQ scan. The diagnosis was established in case of an abnormal angiogram or a high-probability VQ scan. The SimpliRed D-dimer test and CUS were performed upon inclusion in the study and a spiral CT was only performed in patients with an abnormal VQ scan. Test characteristics were calculated for subgroups of patients defined by their VQ scan result.

Results The overall likelihood ratio of a normal SimpliRED test was 0.3 (95% CI 0.2-0.5). In the subgroup of 115 patients with a non-diagnostic VQ scan the likelihood ratio was significantly higher 1.0 (95% CI 0.5-1.8). The overall likelihood ratio of an abnormal CUS was 9.3 (95% CI 4.4-19.6). The likelihood ratio decreased to 1.8 (95% CI 0.2-16.5) in patients with a non-diagnostic VQ scan. The likelihood ratio of an abnormal spiral CT in all patients with an abnormal VQ scan was 4.5 (95% CI 2.8-7.1) and the subgroup estimate was 3.1 (95% CI 1.5-6.2). The differences between the subgroups were significant for the SimpliRed D-Dimer when tested with a Breslow-Day test for homogeneity (p=0.03).

Conclusions Although the overall performance of the evaluated non-invasive diagnostic tests seem promising, their discriminatory power in the subgroup of patients with non-diagnostic VQ scans was disappointing. Ignoring these dependencies in decision analyses can result in erroneous recommendations.
Introduction
The diagnosis of pulmonary embolism (PE) is difficult to establish based on clinical signs and symptoms alone, as revealed by the introduction of objective tests in the 1970's. It was shown that only one third of the patients with a clinical suspicion actually had the diagnosis PE. The value of the ventilation-perfusion lung scintigraphy (VQ scan) for the diagnostic work-up of patients suspected of PE was established in 1990's. Nevertheless in approximately 50% of patients the results of a VQ scan are inconclusive and referred to as non-diagnostic. In these patients further objective tests are needed to confirm or refute the diagnosis. The ideal test for this purpose is pulmonary angiography, which is considered the reference test for PE. However, pulmonary angiography is invasive, time consuming and associated with morbidity.

Several non-invasive tests, such as d-dimer assays, compression ultrasonography (CUS) and spiral CT, are advocated to reduce the number of pulmonary angiograms in patients with suspected PE. Most of these recommendations are based on decision analyses in which the clinical effects of the application of diagnostic strategies are modeled. In general, these decision analyses combine reports on test-characteristics of the different tests in a decision tree to calculate the proportion of false-positive and false-negative results of an entire strategy. The test-characteristics in these reports, and thus, those used in these decision analyses, are usually assessed in a group of consecutive patients suspected of PE. However, some studies have shown that the diagnostic accuracy of a test is not constant but can differ across subgroups. Consequently, incorporating subgroup estimates rather than overall estimates can alter the results, and recommendations, of decision analyses.

To assess the validity of the independence assumption in decision analyses, we examined the diagnostic accuracy of non-invasive tests in all patients suspected of PE and in subgroups of patients, as determined by their VQ scan result, in a recent multicenter study on the accuracy of a D-dimer test, CUS and spiral CT.

Methods

Patients
Data were obtained in a large-scale study in six Dutch teaching hospitals. Previous reports of this trial include the anatomic distribution of pulmonary embolism, the diagnostic accuracy of a fibrin degradation product assay, clinical decision rules and spiral CT. From May 1997 through March 1998 both in- and outpatients with a clinical suspicion of PE were eligible for the study. Patients were excluded if they
were younger than 18 years of age, were pregnant, had an indication for acute thrombolytic therapy, had already undergone objective diagnostic testing for venous thromboembolism or if there was an inability to complete the diagnostic protocol within 48 hours of presentation. The study protocol was approved by the Institutional Review Boards of all hospitals and written informed consent was obtained from all participating patients.

**Diagnostic Investigations**

A detailed clinical history, physical examination, D-dimer test, CUS and a VQ scan were performed within 24 hours of presentation in all patients. Spiral CT angiography was only performed in patients with an abnormal VQ scan. Pulmonary angiography was performed in case of a non-diagnostic VQ scan and whenever a high probability VQ scan was followed by a normal spiral CT scan. The study protocol demanded that all diagnostic investigations were performed within 48 hours of presentation. Patients were classified as having PE on the basis of a high probability VQ scan or abnormal angiogram. PE was excluded by a normal VQ scan or a normal pulmonary angiogram.

Upon study inclusion, and prior to or within 24 hours of the start of heparin therapy, the rapid whole blood SimpliRED D-dimer test (Agen Biomedical Ltd, Brisbane, Australia) was performed. The test was done with capillary blood (2x10^6/l) drawn by fingerstick. The test is abnormal, when agglutination of red cells is observed within two minutes, which indicates a plasma concentration of D-dimer fragments of 0.20 mg/l or above. At each centre, the test was performed by a limited number of investigators who had previously been trained in the interpretation of the SimpliRED. The test result was recorded without knowledge of the outcome of other diagnostic tests.

CUS was performed by experienced examiners unaware of the outcome of the other diagnostic investigations. A two-point B-mode CUS was performed using a 5-10 MHz linear array transducer. With the patient in the supine position, the common femoral vein was located in the groin, using the adjacent artery as a reference point. The popliteal vein was subsequently located in the mid-popliteal fossa with the patient in the prone position. All veins were imaged in the transverse plane to assess their compressibility. Ultrasonography was considered abnormal if a venous segment could not be completely compressed, as previously described.13

Spiral CT angiography was performed during a 32-second single breath hold. If patients were very dyspnoeic, scanning was performed during shallow breathing. Images were reconstructed at 2mm intervals and interpreted on a viewing station. All spiral CT scans were interpreted independently by two experienced examiners.
In case of disagreement, the interpretation of a third examiner was considered decisive. PE was considered to be present if in case of a well opacified scan there was an intraluminal filling defect on more than one slice and no artefacts were present. A filling defect could be seen as complete occlusion of the vessel, an eccentric partial filling defect or a partial central filling defect surrounded by contrast agent. A CT-scan was considered negative if in case of a good quality scan no filling defects could be seen.\(^\text{14}\)

Six-view perfusion lung scintigraphy was performed using 100 Mbq of \(^{99m}\)Technetium-labelled macro-aggregates of albumin. Ventilation scintigraphy was performed using \(^{81m}\)Kryptongas in case of segmental or larger perfusion defects. The VQ scans scans were read by two nuclear physicians in consensus. The VQ scans were interpreted by using a lung segment reference chart, and were classified as being normal (no perfusion defects), high probability (at least one segmental or larger perfusion defects with locally normal ventilation) or non-diagnostic (ventilation-perfusion defects not fulfilling the criteria for high probability or normal).\(^\text{4,15}\)

Pulmonary Angiography was performed using a digital subtraction technique, with the catheter positioned selectively in the left and right pulmonary artery. Images were obtained in at least two projections. All pulmonary angiograms were interpreted independently by two experienced examiners. In case of disagreement, the interpretation of a third examiner was considered decisive. The angiograms were considered normal or abnormal according to previously described criteria.\(^\text{16}\)

**Analysis**

The analysis was restricted to the patients in whom the diagnosis of PE was established or refuted as defined in the protocol. Likelihood ratios were calculated for normal and abnormal test results of the SimpliRed D-dimer, CUS and the spiral CT as measure of diagnostic accuracy. To examine the value of the different tests in patients with a non-diagnostic VQ scan, the analysis was repeated in subgroups defined by the VQ scan results (high probability, non-diagnostic or normal). For each likelihood ratio 95% confidence intervals were calculated according to the normal approximation of the binomial distribution. To compare the results within the subgroups we calculated the Diagnostic Odds Ratio (DOR). The DOR is a single measure to describe the diagnostic accuracy of a test and can be calculated as \(\text{likelihood ratio of a positive test divided by the likelihood ratio of a negative test result.}\)\(^\text{17}\) The DOR ranges from 0 to infinity, with higher values indicating better test performance and unity indicating no discrimination between diseased and non diseased. The differences in diagnostic accuracy of a test within the three subgroups were examined with the Breslow-Day test for homogeneity of odds ratios.\(^\text{18}\)
value \leq 0.05 was considered significant to refute the null-hypothesis that the diagnostic accuracy between the subgroups was the same.

Results

A total of 1162 consecutive patients with clinically suspected PE were screened. Of these patients, 179 were excluded for the following reasons: need for acute thrombolytic therapy (n=5), pregnancy (n=11), age under 18 years (n=16), objective diagnostic tests for PE already performed prior to study entry (n=43) and an expected inability to complete the diagnostic work-up within 48 hours (n=104). Hence, a total of 983 patients were eligible for inclusion in the study of whom 627 (64%) gave informed consent. A final diagnosis regarding the presence or absence of PE was not reached in 110 of the 627 included patients because of withdrawal of informed consent, clear evidence for an alternative diagnosis, medical reasons or technical failure. The baseline characteristics of the 517 study patients in whom a final diagnosis was obtained according to protocol were similar to those of the initially included 627 patients.

A SimpliRED D-dimer and a CUS with a subsequent VQ scan test result was available in 490 and 471 patients respectively. In 8 patients the results of the VQ scan were inconclusive. In an additional 19 and 38 patients respectively the SimpliRED test and the CUS results were not available. The Spiral CT result was available in 230 of the 274 patients in which it was indicated according to the protocol. Spiral CT was not performed in 36 patients due to logistic reasons and had inconclusive results in 8 patients.

Table 1. Clinical and demographic characteristics of the 517 study patients with clinically suspected PE, in whom the final diagnosis was obtained as well as the 627 initially included patients.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Study patients (n=517)</th>
<th>Initially included patients (n=627)</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>215 (42%)</td>
<td>270 (43%)</td>
</tr>
<tr>
<td>mean age, years (SD)</td>
<td>51 (18)</td>
<td>53 (18)</td>
</tr>
<tr>
<td>out-patients</td>
<td>417 (81%)</td>
<td>490 (78%)</td>
</tr>
<tr>
<td>median duration of symptoms, days (quartiles)</td>
<td>3 (1,9)</td>
<td>3 (1,9)</td>
</tr>
<tr>
<td>previous history of VTE</td>
<td>73 (14%)</td>
<td>98 (16%)</td>
</tr>
<tr>
<td>family history of VTE</td>
<td>105 (20%)</td>
<td>122 (19%)</td>
</tr>
<tr>
<td>risk-period*</td>
<td>193 (37%)</td>
<td>249 (40%)</td>
</tr>
<tr>
<td>active malignancy</td>
<td>50 (10%)</td>
<td>71 (11%)</td>
</tr>
<tr>
<td>symptoms of DVT</td>
<td>31 (6%)</td>
<td>43 (7%)</td>
</tr>
</tbody>
</table>

VTE = venous thromboembolism, * = period of immobilization, surgery or trauma in period of 3 months before presentation, DVT = deep vein thrombosis
The performance of the SimpliRED D-dimer assay, CUS and spiral CT in the total group and in subgroups defined by the VQ scan results are shown in table 2. The overall likelihood ratio of a normal SimpliRED test result was 0.3. In the subgroups defined by the results of the VQ scan the discriminative value of the test diverged. The lowest likelihood ratio of a normal test was observed in the subgroup of patients with a normal VQ scan. Whereas in the 115 patients with a non-diagnostic VQ scan the likelihood ratio was 1.0, showing no discriminatory potential of the test for the presence or absence of PE. A similar trend towards unity for patients with a non-diagnostic VQ scan was observed for the likelihood ratio of a abnormal D-dimer result. The discriminatory performance of the SimpliRED assay was significantly different across the subgroups (Breslow-Day test; p=0.03).

The overall likelihood ratio of an abnormal CUS was 9.3. In the subgroups defined by the results of the VQ scan the highest likelihood ratio of an abnormal test, 9.0, was observed in the subgroup of patients with a high probability VQ scan. In the 116 patients with a non-diagnostic VQ scan the likelihood ratio declined to 1.8. The likelihood ratio of a normal CUS deteriorated further from 0.8 to 1.0, no discriminatory value, in the subgroups of patients with a non-diagnostic or a normal VQ scan. However, discriminatory performance measured with the DOR of the

Table 2. Diagnostic accuracy of Simpli-Red D-dimer, Compression ultrasonography and Spiral CT in all patients suspected of PE and in subgroups of patients, as determined by their VQ scan result

<table>
<thead>
<tr>
<th>Tests</th>
<th>N</th>
<th>% PE</th>
<th>LR+ (95% CI)</th>
<th>LR- (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D-Dimer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total group</td>
<td>490</td>
<td>31</td>
<td>2.1 (1.8-2.4)</td>
<td>0.3 (0.2-0.5)</td>
</tr>
<tr>
<td>high probability</td>
<td>147</td>
<td>89</td>
<td>1.6 (1.0-2.7)</td>
<td>0.4 (0.2-0.7)</td>
</tr>
<tr>
<td>non-diagnostic</td>
<td>115</td>
<td>16</td>
<td>1.0 (0.7-1.5)</td>
<td>1.0 (0.5-1.8)</td>
</tr>
<tr>
<td>normal</td>
<td>228</td>
<td>1</td>
<td>3.6 (2.9-4.5)</td>
<td>0.2 (0.0-2.3)*</td>
</tr>
<tr>
<td><strong>CUS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total group</td>
<td>471</td>
<td>32</td>
<td>9.3 (4.4-19.6)</td>
<td>0.8 (0.7-0.9)</td>
</tr>
<tr>
<td>high probability</td>
<td>145</td>
<td>89</td>
<td>9.0 (0.6-141)*</td>
<td>0.7 (0.7-0.8)</td>
</tr>
<tr>
<td>non-diagnostic</td>
<td>116</td>
<td>16</td>
<td>1.8 (0.2-16.5)</td>
<td>1.0 (0.9-1.1)</td>
</tr>
<tr>
<td>normal</td>
<td>210</td>
<td>1</td>
<td>4.7 (0.6-71.8)*</td>
<td>1.0 (1.0-1.0)</td>
</tr>
<tr>
<td><strong>SPCT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total group</td>
<td>230-</td>
<td>54</td>
<td>4.5 (2.8-7.1)</td>
<td>0.4 (0.3-0.5)</td>
</tr>
<tr>
<td>high probability</td>
<td>121</td>
<td>89</td>
<td>4.6 (1.3-16.7)</td>
<td>0.3 (0.2-0.5)</td>
</tr>
<tr>
<td>non-diagnostic</td>
<td>109</td>
<td>16</td>
<td>3.1 (1.5-6.2)</td>
<td>0.6 (0.4-1.0)</td>
</tr>
</tbody>
</table>

* added 0.5 to all cells of the 2x2 contingency table to estimate LR, ~ Small group as a spiral CT (SPCT) was not performed in patients with a normal VQ scan. CUS is compression ultrasonography.
subgroups was not significantly different when tested with the Breslow-Day test (p=0.19).

The diagnosis PE was confirmed in 125 of the 230 (54%) of the patients in whom a spiral CT was performed. The likelihood ratio of an abnormal spiral CT in all patients with an abnormal VQ scan was 4.5. In the subgroup with a non-diagnostic VQ scan the likelihood ratio of an abnormal test declined to 3.1. The poorest discriminative value of a negative test result was also observed in this subgroup. In spite of this the DOR of the subgroups of patients with a high probability VQ scan was not significantly from the patients with a non-diagnostic VQ scan (p=0.28).

Discussion

Our analysis showed that the diagnostic accuracy of non-invasive tests for PE differs in subgroups of patients defined by their VQ scan result. We observed a significant poorer diagnostic performance of the SimpliRED assay in patients with a non-diagnostic VQ scan. The trend towards a less informative test in this subgroup was also observed for the spiral CT as well as for the CUS.

Unlike the predictive value of a diagnostic test, which depends on the prevalence of disease, likelihood ratios, sensitivity and specificity are often assumed to be constant. Interestingly, in 1984 Hlatky et al. already showed that sensitivity and specificity can differ significantly between subgroups of patients defined by clinical characteristics using exercise electrocardiography as an example. Recently Moons et al confirmed these findings examining the effect of variables from history and physical examination on the diagnostic performance of the same test. In this study we demonstrated that variation of test characteristics also depend on previous test results. This could be expected on basis of the previous mentioned studies as theoretically there is no difference between a history item and a diagnostic test. Both can be seen as a test with a potential to discriminate groups of patients into high and low-risk for the disease under study.

The biological explanation for these differences is the fact that a disease status is seldom a true dichotomous classification. Often there is a spectrum of disease ranging from small limited forms sometimes even without symptoms to more extensive conditions with severe symptoms. The test characteristics can differ for patients with severe symptoms compared to patients with limited symptoms, as was demonstrated by Lachs and colleagues. If patients have been selected based on previous test results, the spectrum of disease in the remaining patients may be limited. Usually patients with very severe forms of disease can be identified easily. For example a PE can be large, obstructing a whole lobe and causing life-threatening disease or small and isolated, located in a subsegmental artery causing minor
complaints. In our study, the large obvious PE and completely normal patients have already been identified by the VQ scan, resulting in a subgroup of patients with a limited spectrum of the disease which are more difficult to classify. Consequently a generalization of the ‘unconditional’ likelihood ratios of a positive or negative test of the SimpliRED D-dimer assay will overestimate its diagnostic performance when it is applied after a non-diagnostic VQ scan.

Our results could potentially suffer from a selection bias as not in all patients a final diagnosis was obtained and thus could not be accounted for in the final analysis. However, the baseline characteristics of the final study population were similar to those of the initially included patients. One could argue that the differences we have found are a coincidence due to statistical variation and not applicable to all patients with a non-diagnostic VQ scan. Although the number of patients in the subgroups is small, we do believe our results are generalisable as the trend towards a less informative test in the subgroup of patients with a non-diagnostic VQ scan is observed for all three tests examined and there is a plausible biological explanation.

The variations of test characteristics reported here can have clinical implications, as many currently used diagnostic strategies are based on decision analysis using overall estimates of test characteristics instead of subgroup estimates. For example, two recent cost-effectiveness analyses that compared different diagnostic strategies for the diagnosis of PE used 0.09 and 0.07 as estimates for the likelihood ratio of a negative test, based on reports of the overall test-characteristics. (6, 21). Our results suggest that these values are considerable lower if a D-dimer test is applied in patients with a non-diagnostic VQ scan. Thus, the calculations in these studies for a strategy with a D-dimer test after a VQ scan are probably overoptimistic. This can have important consequences, since diagnostic strategies with this test sequence can erroneously dominate competing strategies, leading to incorrect recommendations. The optimal solution to overcome this problem is to use test characteristics conditional on all prior information for each test in a decision tree. However, such data is often lacking in current reports of diagnostic test evaluations. An alternative is to downgrade the overall test characteristics of a test when a test is applied after other tests in a strategy and/or to examine a range of plausible values in a sensitivity analysis.

In conclusion, although the overall performance of non-invasive diagnostic tests their discriminatory power in the subgroup of patients with non-diagnostic VQ scans is disappointing. Ignoring these dependencies in decision analyses comparing diagnostic strategies can result in overoptimistic estimates for non-invasive strategies and thus result in erroneous recommendations.
Acknowledgements
The authors thank Prof. Patrick Bossuyt for his suggestions to the manuscript and in specific the title

References


