Comparison of Observer Variability and Accuracy of Different Criteria for Lung Scan Interpretation

Petronella J. Hagen, MD[1,2]; Ieneke J.C. Hartmann, PhD[3,4]; Otto S. Hoekstra, PhD[2,5]; Marcel P.M. Stokkel, PhD[3]; Pieter E. Postmus, PhD[1]; and Martin H. Prins, PhD[6]; for the ANTELOPE Study Group

[1]Department of Pulmonary Medicine, Vrije Universiteit Medical Center, Amsterdam, The Netherlands; [2]Department of Nuclear Medicine, Vrije Universiteit Medical Center, Amsterdam, The Netherlands; [3]Department of Nuclear Medicine, University Medical Center, Utrecht, The Netherlands; [4]Department of Radiology, University Medical Center, Utrecht, The Netherlands; [5]Department of Clinical Epidemiology, Vrije Universiteit Medical Center, Amsterdam, The Netherlands; and [6]Department of Clinical Epidemiology and Medical Technology Assessment, Academic Hospital Maastricht, Maastricht, The Netherlands

Different criteria have been advocated for the interpretation of ventilation/perfusion (V/Q) lung scans in patients with suspected pulmonary embolism (PE). Besides these predefined criteria, many physicians use an integration of the different sets of criteria and their own experience—the so-called Gestalt interpretation. The purpose of this study was to evaluate interobserver variability and accuracy of 3 sets of criteria: the Hull and PIOPED (Prospective Investigation of Pulmonary Embolism Diagnosis) criteria and the Gestalt interpretation. **Methods:** Two experienced observers interpreted V/Q scans of all 328 patients according to the 3 different schemes. The diagnostic classification obtained for the different sets of criteria was analyzed against the presence or absence of PE. **Results:** The interobserver variabilities as assessed by the κ statistics of the PIOPED and Hull criteria and for the Gestalt interpretation were 0.70 (95% confidence interval [CI], 0.64–0.76), 0.79 (95% CI, 0.73–0.85), and 0.65 (95% CI, 0.58–0.72), respectively. The differences in κ values between the Hull and PIOPED criteria and between the Hull criteria and Gestalt interpretation were statistically significant ($P < 0.05$ and $P < 0.001$, respectively). For 16 patients (14 without PE) with a normal lung scan result according to the Hull criteria, the result according to the PIOPED criteria was low probability. For 21 patients (12 with PE), the scans were intermediate probability according to the PIOPED criteria, whereas the result with the Hull criteria was high probability. Analysis of receiver-operating-characteristic curves yielded a comparable area under the curve for all sets of criteria (0.87–0.90). **Conclusion:** The Hull, PIOPED, and Gestalt interpretation of V/Q lung scans all have a good accuracy and interobserver variability. However, the reproducibility of the Hull criteria is superior in comparison with that of the other sets of criteria.

Key Words: ventilation/perfusion lung scintigraphy; pulmonary embolism; criteria

J Nucl Med 2003; 44:739–744

Pulmonary embolism (PE) remains a diagnostic challenge, with 2 opposite aims: to effectively confirm or to safely exclude the presence of venous thromboembolism. Objective testing is necessary to establish a final diagnosis because only approximately 25% (*1,2*) of the patients indeed have PE. Although pulmonary angiography is the gold standard for diagnosing PE, it is usually not considered as the test of first choice in the diagnostic work-up of patients with suspected PE because of its invasive nature and limited availability. Ventilation/perfusion (V/Q) lung scanning is primarily used as the pivotal test. Lung scanning is much less invasive than pulmonary angiography, is associated with negligible morbidity, and can be performed in most hospitals with nuclear medicine facilities.

Several different diagnostic classification schemes have been suggested for the interpretation of a V/Q scan. An ideal set of criteria would minimize the number of nonconclusive scan results and have both high positive and negative predictive values. A lung scan result can be classified as normal (or near normal), nondiagnostic (low or intermediate probability) for PE, or high probability. Each of these categories has clinical implications. A normal perfusion scan rules out clinically important PE and, therefore, anticoagulant therapy can be safely withheld (*3*). A high-probability lung scan result confirms the diagnosis PE (positive predictive value, approximately 90%) and justifies treatment with anticoagulants (*2*). In patients with a nondiagnostic lung scan, further investigations are required to confirm or refute the diagnosis. Although the accuracy and reproducibility have been issues of controversy, the (revised) PIOPED (Prospective Investigation of Pulmonary Embolism Diagnosis) criteria (*4*) and the Hull criteria (*1*) are the most frequently used interpretations for V/Q lung scan readings. The major differences in both sets of criteria are shown in Table 1. However in daily practice many nuclear medicine physicians do not adhere rigidly to a single diagnostic scheme but

TABLE 1
Main Differences in (Revised) PIOPED Criteria and Hull Criteria

PIOPED criteria	Hull criteria
High probability (≥2 segments)	High probability (≥1 segment)
Intermediate probability	Indeterminate probability
Low probability	
Normal perfusion	Normal perfusion

Revised PIOPED criteria (3).
Hull criteria (4).

appear to use an integration of the different sets of criteria and his or her own experience—the so-called Gestalt interpretation (5). Because of differences in these methods of interpretation, the patient's management may vary, with consequences for further investigations, treatment, and costs.

The purpose of this study was to evaluate the performance of different observers using 3 sets of criteria for the interpretation of the V/Q scan in patients with suspected PE: the Hull and (revised) PIOPED criteria and the Gestalt interpretation.

MATERIALS AND METHODS

In each of 3 clinical centers, all patients with suspected PE and for whom a request for V/Q scan was made between May 1997 and March 1998 were considered for study entry. This study was part of a larger multicenter trial on diagnostic methods in suspected PE (6–10). The eligible study population consisted of both in- and outpatients, ≥18 y old, who were not pregnant, did not have an indication for thrombolytic therapy, and in whom objective examinations for diagnosing venous thromboembolic disease according to patient's current symptoms had not already been made. The study was designed as an "intention to diagnose" study, which means the patients with contraindications for spiral CT angiography or conventional pulmonary angiography were not excluded from participating in the study. The Institutional Review Boards of all participating centers approved the study and informed consent was obtained from all patients enrolled in the study.

Study Protocol

Before any further objective testing, the attending physicians were asked to give a probability estimate for PE, based on evaluation of the clinical history, physical examination, chest radiography, and electrocardiography. Within 24 h of referral, the patients underwent venous duplex ultrasonography of the deep leg veins, D-dimer test, and a perfusion scan. Patients were stratified according to the lung scan result. If the perfusion scan was normal, the investigation was stopped and no further tests were performed. Ventilation scintigraphy with [81m]Kr was indicated in all patients with at least 1 segmental perfusion defect. Whenever possible, the ventilation study was performed on the same day, but at least within 24 h after the perfusion scan. When >24 h elapsed between the first perfusion scan and the ventilation scan, a second perfusion scan was obtained. Spiral CT angiography was performed on all patients with perfusion defects, irrespective of the size of these defects. Pulmonary angiography was performed on patients with a non–high-probability V/Q lung scan and on patients with discordance between V/Q lung scan and spiral CT angiography (high-probability V/Q lung scan and a normal spiral CT scan). In patients with a contraindication for spiral CT angiography or pulmonary angiography, the study protocol was violated. However, these patients were not excluded from the study.

We aimed at performing the complete study protocol within 48 h after the first V/Q scintigraphy, with a maximum of 24 h between the examinations under study.

According to the protocol, the diagnosis of PE was made on the basis of pulmonary angiography or spiral CT angiography (the latter only in case of a high-probability V/Q lung scan result). A normal perfusion scan or pulmonary angiography ruled out PE. In all cases, the final diagnosis was established by independent, blinded reading of the diagnostic imaging techniques by a panel of experts.

[81m]Kr V/Q Scintigraphy

Perfusion lung scintigraphy was performed within 24 h after referral using 50 MBq ^{99m}Tc-labeled macroaggregated albumin. The tracer was injected intravenously with the patient in the supine position, whereas imaging was performed in a sitting position. Acquisition was performed in at least 4 standard positions (anterior, posterior, left and right posterior oblique) with at least 150 kilocounts per second per view (low-energy, high-resolution [LEHR] collimator, 128 × 128 matrix). In almost all patients, 6 view images were available.

Ventilation scintigraphy with [81m]Kr gas was performed either immediately after perfusion scintigraphy (medium-energy, high-resolution collimator, 128 × 128 matrix) or using dual-isotope scanning (LEHR collimator, 128 × 128 matrix). In case [81m]Kr was not available, inhalation imaging was executed the next day, but at least within 24 h. Each image was made with at least 200 kilocounts per second per view. Ventilation scans were performed in the same projections as the perfusion scans.

V/Q Scintigraphy Assessment

The V/Q lung scans were interpreted immediately by the nuclear medicine physician on duty, with all clinical information available at that time. In this session the Hull criteria were used for the assessment of the V/Q lung scan result because these criteria are advised in the Dutch consensus for the diagnosis of PE (11). On later occasions, 2 experienced observers read all scans again independent from each other. The readers were unaware of clinical data and other test results. In all sessions a lung segment reference chart was available (12). For those patients in whom perfusion scintigraphy was repeated another day and perfusion defects differed, the first perfusion scan was used for the V/Q lung scan assessment. In each session, scans were interpreted only according to 1 set of criteria (i.e., Hull, PIOPED, or Gestalt). For the Gestalt interpretation the observers were asked to make a probability estimate of the presence of PE on a visual analog scale of 0%–100%. The estimate was based solely on the personal experience and opinion of the observers. Before analysis, categories of <20%, 20%–80%, and >80% were defined.

The diagnostic classification obtained for the different sets of criteria was compared with the final diagnosis according to the protocol.

Statistics

The agreement between the Hull and PIOPED criteria and the Gestalt interpretation and the interobserver variability corrected for chance were evaluated with κ statistics. A κ value of 1 corresponds to perfect agreement; 0 corresponds to agreement as expected by chance (*13*).

Receiver-operating-characteristic (ROC) analysis and areas under the ROC curve (AUC) were used as objective measures to evaluate the overall accuracy of the Hull and PIOPED criteria and the Gestalt interpretation (*14,15*).

Statistical analyses were performed with SPSS statistical software (SPSS, Inc., Chicago, IL).

RESULTS

Patient Population

During the course of the study, 693 consecutive patients were referred for clinically suspected PE. Of these, 107 were excluded on the basis of the predefined exclusion criteria (8 were pregnant, 10 were <18 y old, 3 had an indication for thrombolytic therapy, 28 already had diagnostic tests performed, and in 58 patients it was expected that the protocol could not be completed within 48 h or patients were unable to give informed consent). Of the remaining 586 eligible patients, 328 (56%) provided informed consent and underwent perfusion scintigraphy. The demographic and clinical characteristics of these patients are detailed in Table 2.

Availability of Ventilation Scintigraphy

In 316 of the 328 patients, a definitive conclusion about the presence or absence of PE was available. On the basis of this conclusion, 117 patients had a normal perfusion lung scan, 54 had subsegmental perfusion defects, and 145 had at least 1 segmental perfusion defect and, therefore, an indication for ventilation scintigraphy.

TABLE 2
Demographic and Clinical Characteristics of 328 Study Patients with Clinically Suspected PE

Characteristic	Study patients* (n = 328)
Female	187 (57%)
Mean age ± SD (y)	53.3 ± 17.4
Inpatient	77 (23%)
Median duration of symptoms before investigation (d)	2 (1–8)†
History of PE/DVT	56/325 (17%)
COPD	51/328 (16%)
Malignancy	38/296 (13%)
Recent surgery	60/327 (18%)

*Numerator varies because data of some patients were missing.
†Interquartile range.
PE/DVT = PE/deep vein thrombosis; COPD = chronic obstructive pulmonary disease.

TABLE 3
Interobserver Variability for V/Q Lung Scan Assessment Using PIOPED Criteria

Observer 1	Observer 2				
	Normal	Low	Intermediate	High	Total
Normal	94	14			108
Low	15	48	28	1	92
Intermediate	1	4	68	4	77
High				5	46	51
Total	110	66	101	51	328

Interobserver Agreement Between 3 Sets of Criteria

The results of the 2 readers for the different sets of criteria are shown in Tables 3–5. Disagreement between the PIOPED criteria in 2 or more categories—1 assigning a normal result and the other assigning an intermediate- or high-probability pattern—occurred in 2 of the 328 patients (0.6%). The calculated κ values of the PIOPED and Hull criteria and for the Gestalt interpretation were 0.70 (95% confidence interval [CI], 0.64–0.76), 0.79 (95% CI, 0.73–0.85), and 0.65 (95% CI, 0.58–0.72), respectively. The differences in κ values between the Hull and PIOPED criteria and the Hull criteria and the Gestalt interpretation were statistically significant (*P* < 0.05 and *P* < 0.001, respectively).

The association between the classification according to the Hull and PIOPED criteria for observer 1 is shown in Table 6 (the results of observer 2 were similar; data not shown). In 16 patients with a normal lung scan according to the Hull criteria, the PIOPED result was low probability, which would likely result in additional investigations. The final diagnosis in 14 of these 16 patients was no PE, whereas in the 2 other patients no final diagnosis could be achieved. Of the 73 patients with a high-probability V/Q lung scan result according to the Hull criteria, 21 had an intermediate result according to the PIOPED criteria and, thus, further investigations would be required. The final diagnosis of 17 of these patients was available: 12 had PE, 5 had no embolism.

Accuracy of Different Criteria

ROC curves for the 3 sets of criteria were calculated for both observers and are shown in Figures 1 and 2. For this

TABLE 4
Interobserver Variability for V/Q Lung Scan Assessment Using Hull Criteria

Observer 1	Observer 2			
	Normal	Indeterminate	High	Total
Normal	102	17		119
Indeterminate	11	117	8	136
High		9	64	73
Total	113	143	72	328

TABLE 5
Interobserver Variability for V/Q Lung Scan Assessment Using Gestalt Interpretation

Observer 1	Observer 2			
	<20%	20%–80%	>80%	Total
<20%	174	14	3	191
20%–80%	26	39	17	82
>80%		5	50	55
Total	200	58	70	328

analysis the final diagnosis according to the protocol was available in 265 of the 328 patients. The AUCs for the PIOPED and Hull criteria and for the Gestalt interpretation were 0.90 (95% CI, 0.85–0.94), 0.89 (95% CI, 0.85–0.94), and 0.89 (95% CI, 0.85–0.94), respectively, for observer 1. The AUCs of observer 2 were similar: PIOPED, 0.87 (95% CI, 0.83–0.92); Hull, 0.87 (95% CI, 0.82–0.92); and Gestalt, 0.88 (95% CI, 0.83–0.93).

DISCUSSION

In this study 3 different widely used schemes for interpreting V/Q lung scans were compared in consecutive patients with clinically suspected PE. All schemes had a good interobserver variation (0.65–0.79). Nevertheless, the interobserver variability was statistically significantly better when V/Q lung scans were interpreted according to the Hull criteria. The accuracy of the different interpretation schemes, as assessed by ROC curve analysis, was similar and the AUCs varied between 0.87 and 0.90. However, use of the PIOPED criteria rather than the Hull criteria would lead to an increase of additional investigations needed for a final diagnosis in patients with suspected PE.

In several earlier, but smaller, studies, different interpretation schemes for V/Q lung scans were also compared. One study compared the PIOPED criteria with 2 other proposed sets of criteria (Biello and McNeil) in 96 patients with suspected PE, who also underwent pulmonary angiography (16). The study failed to demonstrate statistically significant differences between the different sets of criteria. The best area under the ROC curve observed was 0.85, which is similar to what was found in our study.

In a study by Sostman et al. (4), a group of 105 patients with suspected PE was evaluated comparing the revised

TABLE 6
Comparison of V/Q Lung Scan Classification with Hull vs. PIOPED Criteria Assessed by Observer 1

Hull criteria	PIOPED criteria				
	Normal	Low	Intermediate	High	Total
Normal	103	16			119
Indeterminate	5	73	56	2	136
High		3	21	49	73
Total	108	92	77	51	328
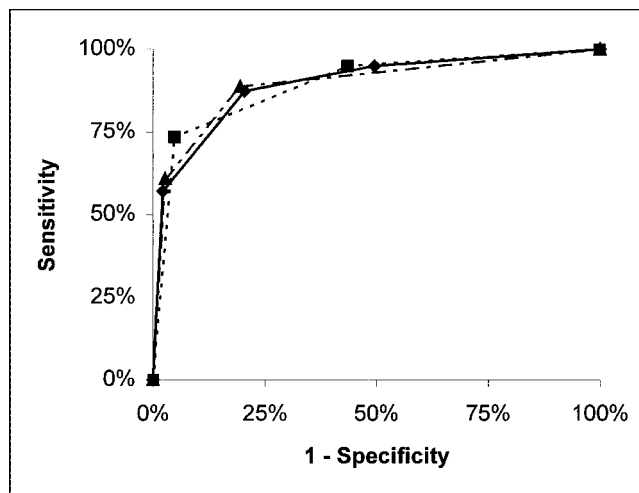
FIGURE 1. ROC curves of PIOPED criteria (♦), Hull criteria (■), and Gestalt V/Q lung scan interpretation (▲), observer 1.

PIOPED criteria with the original PIOPED criteria and a percentage probability based on the reader's own individual experience and subjective impression of the likelihood of PE (Gestalt interpretation). This Gestalt interpretation was the most accurate (area under ROC curve, 0.84). The revised PIOPED criteria were more accurate than the original PIOPED criteria (area under ROC curve, 0.75 vs. 0.65). However, these results were based on consensus readings. For the individual readers, the areas under the ROC curve varied from 0.78 to 0.83 for the Gestalt interpretation and from 0.68 to 0.76 for the revised PIOPED criteria.

In another study, the Gestalt interpretation was compared with the McNeil and Biello schemes in 98 patients with suspected PE. Again, no significant differences were found in overall accuracy (5).
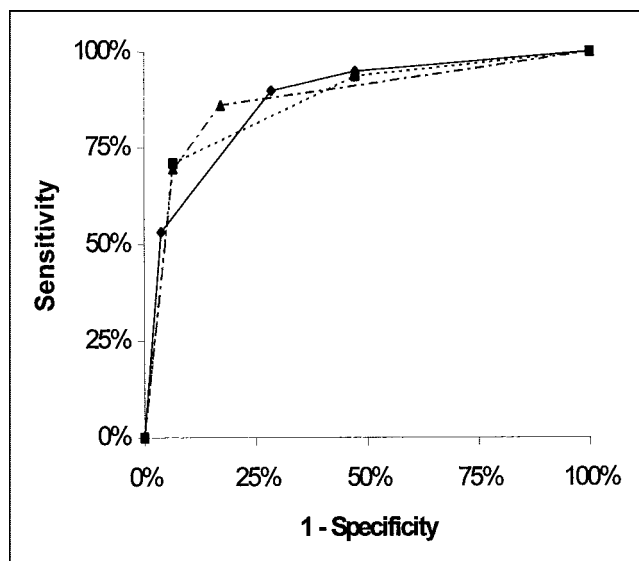
FIGURE 2. ROC curves of PIOPED criteria (♦), Hull criteria (■), and Gestalt V/Q lung scan interpretation (▲), observer 2.

Christiansen et al. (*17*) found moderate and fair κ values (0.31–0.54) comparing different observers using the revised PIOPED criteria. However, all observers showed good accuracy when their scintigraphic diagnoses were compared with that based on angiography. The area under the ROC curves was in the range 0.81–0.89.

In a case control study of 87 patients with asymptomatic PE and 50 patients with symptomatic PE, Kraemmer et al. (*18*) used far more simple criteria to classify the lung scans. The observers were asked to classify the scans as normal, PE (perfusion/ventilation mismatch), or parenchymal lung disease (matched perfusion/ventilation defect). The κ values for the interobserver variation varied from 0.66 to 0.72.

In comparison with these literature observations, the interobserver variation in our study seems in the upper range, whereas the accuracy of the Gestalt interpretation was similar and that of the PIOPED criteria was better than reported previously. This is further illustrated by the number of disagreements of 2 or more categories between the 2 observers, which occurred in our study in only 2 of the 328 patients (0.6%). In 2 studies of Christiansen et al. (*17,19*), disagreement in 2 or more categories was observed in 21 of the 170 patients (12%) and 5 of the 192 patients (3%), respectively. This might be related to the experience of the 2 observers but could also be related to the systematic availability of the lung segment reference chart (*12*).

On the basis of the difference in area under the ROC curves between the revised PIOPED criteria and the Gestalt interpretation, it was suggested that further refinement of these PIOPED criteria could be possible (*4*). However, the results of our study do not support this finding. On the contrary, our findings indicate that the far simpler Hull criteria, consisting of only 3 different categories, are comparable in accuracy and superior in reproducibility.

We believe that our observations on the accuracy and reproducibility of the revised PIOPED and Hull criteria reflect truly the inherent properties of these classifications because the observers were experienced nuclear physicians who had had several consensus training sessions. Also, in comparison with results reported earlier, the Gestalt interpretation was performing well and the lack of a better performance of the Gestalt interpretation in comparison with the PIOPED and Hull criteria is, therefore, unlikely to be due to inexperience. Moreover, the observers were unaware of each other's results and the results of other tests and, to avoid recall bias, the scans were assessed in a random order with a time interval of at least 4 wk, which varied in practice mostly from 6 wk to 3 mo.

CONCLUSION

In this study the interobserver variation of the Hull and PIOPED criteria and theGestalt interpretation of V/Q lung scans was found to be good. However, the reproducibility of the Hull criteria was superior.

APPENDIX

The results of this study are reported on behalf of the ANTELOPE (Advances in New Technologies Evaluating the Localization of Pulmonary Embolism) Study Group of the Dutch prospective multicenter trial on the diagnosis of PE.

Academic Medical Center, Amsterdam, The Netherlands
B.J. Sanson, MD; M.H. Prins, MD; H.R. Büller, MD

Leiden University Medical Center, Leiden, The Netherlands
W. de Monyé, MD; M.V. Huisman, MD; P.M.T. Pattynama, MD

Leyenburgh Hospital, The Hague, The Netherlands
M.J.L. van Strijen, MD; G.J. Kieft, MD

Slotervaart Hospital, Amsterdam, The Netherlands
M.R. Mac Gillavry, MD; D.P.M. Brandjes, MD

Vrije Universiteit Medical Center, Amsterdam, The Netherlands
P.J. Hagen, MD; O.S. Hoekstra, MD; P.E. Postmus, MD

University Medical Center, Utrecht, The Netherlands
I.J.C. Hartmann, MD; J.D. Banga, MD; P.F.G.M. van Waes, MD

ACKNOWLEDGMENT

REFERENCES

1. Hull RD, Hirsh J, Carter CJ, et al. Diagnostic value of ventilation-perfusion lung scanning in patients with suspected pulmonary embolism. *Chest.* 1985;88:819–828.
2. The PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism: results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). *JAMA.* 1990;263:2753–2759.
3. Hull RD, Raskob GE, Coates G, Panju AA. Clinical validity of a normal perfusion lung scan in patients with suspected pulmonary embolism. *Chest.* 1990;97:23–26.
4. Sostman HD, Coleman RE, DeLong DM, Newman GE, Paine S. Evaluation of revised criteria for ventilation-perfusion scintigraphy in patients with suspected pulmonary embolism. *Radiology.* 1994;193:103–107.
5. Sullivan DC, Coleman RE, Mills SR, Ravin CE, Hedlund LW. Lung scan interpretation: effect of different observers and different criteria. *Radiology.* 1983;149:803–807.
6. de Monye W, van Strijen MJ, Huisman MV, Kieft GJ, Pattynama PM. Suspected pulmonary embolism: prevalence and anatomic distribution in 487 consecutive patients—Advances in New Technologies Evaluating the Localisation of Pulmonary Embolism (ANTELOPE) Group. *Radiology.* 2000;215:184–188.
7. Hartmann IJ, Hagen PJ, Melissant CF, Postmus PE, Prins MH. Diagnosing acute pulmonary embolism: effect of chronic obstructive pulmonary disease on the performance of d-dimer testing, ventilation/perfusion scintigraphy, spiral computed tomographic angiography, and conventional angiography. *Am J Respir Crit Care Med.* 2000;162:2232–2237.
8. Mac Gillavry MR, de Monye W, Lijmer JG, et al. Clinical evaluation of a monoclonal antibody-based enzyme immunoassay for fibrin degradation products in patients with clinically suspected pulmonary embolism: ANTELOPE-Study Group. *Thromb Haemost.* 2000;83:892–895.

9. Sanson BJ, Lijmer JG, Mac Gillavry MR, Turkstra F, Prins MH, Buller HR. Comparison of a clinical probability estimate and two clinical models in patients with suspected pulmonary embolism: ANTELOPE-Study Group. *Thromb Haemost.* 2000;83:199–203.

10. Mac Gillavry MR, Sanson BJ, Buller HR, Brandjes DP. Compression ultrasonography of the leg veins in patients with clinically suspected pulmonary embolism: is a more extensive assessment of compressibility useful? ANTELOPE-Study Group. *Thromb Haemost.* 2000;84:973–976.

11. Briet E, van Beek EJ, Oudkerk M. Consensus diagnosis pulmonary embolism [in Dutch]. *Ned Tijdschr Geneeskd.* 1993;137:255–259.

12. Lensing AW, van Beek EJ, Demers C, et al. Ventilation-perfusion lung scanning and the diagnosis of pulmonary embolism: improvement of observer agreement by the use of a lung segment reference chart. *Thromb Haemost.* 1992;68:245–249.

13. Altman DG. *Practical Statistics for Medical Research.* London, U.K.: Chapman and Hall; 1991:404–408.

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

15. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology.* 1983;148:839–843.

16. Webber MM, Gomes AS, Roe D, La Fontaine RL, Hawkins RA. Comparison of Biello, McNeil, and PIOPED criteria for the diagnosis of pulmonary emboli on lung scans. *AJR.* 1990;154:975–981.

17. Christiansen F, Nilsson T, Mare K, Carlsson A. Observer agreement and accuracy for lung scintigraphy: a comparison with pulmonary angiography. *Nucl Med Commun.* 1997;18:112–117.

18. Kraemmer NH, Husted SE, Krusell LR, Charles P, Fasting H, Hansen HH. Observer variation in the interpretation of ventilation-perfusion lung scintigraphy. *Eur J Nucl Med.* 1994;21:103–107.

19. Christiansen F, Andersson T, Rydman H, Qvarner N, Mare K. Rater agreement in lung scintigraphy. *Acta Radiol.* 1996;37:754–758.