Pulmonary Nodules on Multi–Detector Row CT Scans: Performance Comparison of Radiologists and Computer-aided Detection1

Article *in* Radiology · January 2005

DOI: 10.1148/radiol.2341040589 · Source: PubMed

CITATIONS
198

READS
118

11 authors, including:

Geoffrey D Rubin
Duke University Medical Center
343 PUBLICATIONS **12,761** CITATIONS

SEE PROFILE

David S Paik
Stanford University
93 PUBLICATIONS **2,446** CITATIONS

SEE PROFILE

Anthony J Sherbondy
Stanford University
24 PUBLICATIONS **1,094** CITATIONS

SEE PROFILE

Larry Chow
Oregon Health and Science University
27 PUBLICATIONS **1,323** CITATIONS

SEE PROFILE

Pulmonary Nodules on Multi–Detector Row CT Scans: Performance Comparison of Radiologists and Computer-aided Detection[1]

Geoffrey D. Rubin, MD
John K. Lyo, MD[2]
David S. Paik, PhD
Anthony J. Sherbondy, MS
Lawrence C. Chow, MD
Ann N. Leung, MD
Robert Mindelzun, MD
Pamela K. Schraedley-Desmond, PhD
Steven E. Zinck, MD
David P. Naidich, MD
Sandy Napel, PhD

Published online before print
10.1148/radiol.2341040589
Radiology 2005; 234:274–283

Abbreviations:
CAD = computer-aided detection
FP = false-positive
ROC = receiver operating characteristic
SNO = surface normal overlap
TP = true-positive

[1] From the Departments of Radiology (G.D.R., J.K.L., D.S.P., L.C.C., A.N.L., R.M., P.K.S.D., S.E.Z., S.N.) and Electrical Engineering (A.J.S.), Stanford University School of Medicine, 300 Pasteur Dr, S-072, Stanford, CA 94305-5105; and Department of Radiology, New York University School of Medicine, New York, NY (D.P.N.). Received March 31, 2004; revision requested June 8; revision received July 26; accepted August 19. **Address correspondence to** G.D.R. (e-mail: *grubin@stanford.edu*).

Authors stated no financial relationship to disclose.

[2] **Current address:** Department of Radiology, Yale University School of Medicine, New Haven, Conn.

Author contributions:
Guarantors of integrity of entire study, G.D.R., J.K.L., D.S.P.; study concepts and design, G.D.R., D.S.P., S.N.; literature research, G.D.R., D.S.P., P.K.S.D.; clinical studies, S.E.Z.; data acquisition, G.D.R., L.C.C., A.N.L., R.M., D.P.N., A.J.S.; data analysis/interpretation, G.D.R., J.K.L., D.S.P.; statistical analysis, P.K.S.D.; manuscript preparation, G.D.R., J.K.L., D.S.P.; manuscript definition of intellectual content, G.D.R., D.S.P., S.N.; manuscript revision/review, all authors; manuscript editing and final version approval, G.D.R., S.N.

PURPOSE: To compare the performance of radiologists and of a computer-aided detection (CAD) algorithm for pulmonary nodule detection on thin-section thoracic computed tomographic (CT) scans.

MATERIALS AND METHODS: The study was approved by the institutional review board. The requirement of informed consent was waived. Twenty outpatients (age range, 15–91 years; mean, 64 years) were examined with chest CT (multi–detector row scanner, four detector rows, 1.25-mm section thickness, and 0.6-mm interval) for pulmonary nodules. Three radiologists independently analyzed CT scans, recorded the locus of each nodule candidate, and assigned each a confidence score. A CAD algorithm with parameters chosen by using cross validation was applied to the 20 scans. The reference standard was established by two experienced thoracic radiologists in consensus, with blind review of all nodule candidates and free search for additional nodules at a dedicated workstation for three-dimensional image analysis. True-positive (TP) and false-positive (FP) results and confidence levels were used to generate free-response receiver operating characteristic (ROC) plots. Double-reading performance was determined on the basis of TP detections by either reader.

RESULTS: The 20 scans showed 195 noncalcified nodules with a diameter of 3 mm or more (reference reading). Area under the alternative free-response ROC curve was 0.54, 0.48, 0.55, and 0.36 for CAD and readers 1–3, respectively. Differences between reader 3 and CAD and between readers 2 and 3 were significant ($P < .05$); those between CAD and readers 1 and 2 were not significant. Mean sensitivity for individual readings was 50% (range, 41%–60%); double reading resulted in increase to 63% (range, 56%–67%). With CAD used at a threshold allowing only three FP detections per CT scan, mean sensitivity was increased to 76% (range, 73%–78%). CAD complemented individual readers by detecting additional nodules more effectively than did a second reader; CAD-reader weighted κ values were significantly lower than reader-reader weighted κ values (Wilcoxon rank sum test, $P < .05$).

CONCLUSION: With CAD used at a level allowing only three FP detections per CT scan, sensitivity was substantially higher than with conventional double reading.

Supplemental material: *radiology.rsnajnls.org/cgi/content/full/2341040589/DC1*

The successful detection, characterization, and treatment of a myriad of lung diseases, including both primary and metastatic lung cancers, begins with the accurate identifica-

tion of pulmonary nodules. Some of the most common indications for performing chest computed tomography (CT) are clinical signs or symptoms suspicious for cancer, or possible nodules seen with less-specific imaging tests such as chest radiography.

The detection of pulmonary nodules at CT is influenced substantially by the method of image data acquisition. The development of multi–detector row CT technology has made it possible to acquire volumetric data of the lungs with unprecedented spatial resolution during a single breath hold. Although higher spatial resolution, in principle, allows the detection of smaller nodules, one drawback of high-resolution acquisitions is that many more transverse reconstructions are generated than with thick-section techniques. The interpreter must examine up to 10 times the number of images that previously had to be examined. As a result, the efficiency of the interpreter is adversely affected. Furthermore, the increased likelihood of tedium-induced fatigue may adversely affect diagnostic accuracy, particularly because pulmonary lesions are more difficult to discriminate from adjacent normal vascular structures as section thickness diminishes.

In recognition of the important role that CT currently plays in the detection of pulmonary nodules, we believe that there is a critical need to develop methods of CT analysis that ensure accurate, consistent, and efficient diagnoses while facilitating radiologists' ability to capitalize fully on the added spatial resolution available with thoracic multi–detector row CT with a section thickness of 1.5 mm or less.

Double reading by two trained human observers has been shown to improve the detection of both lung cancers and breast cancers on chest radiographs and mammograms, respectively. The paradigm for double reading is based on the "OR" rule, according to which a positive interpretation is assigned to any finding deemed positive by either of two independent readers (1–4). Double independent readings of mammograms result in a 10%–15% increase in breast cancer detection, compared with single readings (1–3,5,6), but they are also associated with an increase of 1%–10% in the false-positive (FP) rate (3,5). In the assessment of chest radiographs for lung cancer, double independent reading performed according to the "OR" rule results in a 3%–30% (mean, 13%) increase in sensitivity, with a 1%–9% (mean, 5%) decrease in specificity (4).

The increased cost of interpretation when two readers are employed in double reading has motivated the development of computer-aided detection (CAD) methods that could replace the second reader. The use of CAD as a second reader to identify opacities that the radiologist might have missed has been shown to result in a significant increase in sensitivity in the interpretation of mammograms (7) and chest radiographs (8). The latter application of CAD was found to result in a 13%–16% increase in sensitivity without an increase in FP frequency (8). These data support the expectation that radiologists effectively should be able to filter out FP results presented by CAD, thereby increasing the number of true-positive (TP) detections and, therefore, the sensitivity of the method, without a substantial decrease in specificity. This is precisely the role we propose for CAD in lung cancer detection.

Thus, the purpose of our study was to compare the performance of radiologists and our CAD algorithm for pulmonary nodule detection on thin-section thoracic CT scans.

MATERIALS AND METHODS

CT Scan Acquisition

Thoracic CT scans from 21 consecutive outpatients who were referred because of clinical suspicion of pulmonary nodules were retrospectively collected between October 2000 and January 2001. The outpatient group included 16 male and five female patients aged 15–91 years (mean, 64 years). Three patients had extrathoracic malignancy, and one of the three had known metastases. The remainder had or were suspected of having at least one of the following: cardiac disease (coronary artery disease or cardiomyopathy), chronic obstructive pulmonary disease, and inflammatory lung disease. A CT scan from an 82-year-old man with active diffuse mycobacterial infection was excluded from the cohort because of the presence of large areas of heterogeneous parenchymal consolidation and mucus plugs, which hindered the confident identification of pulmonary nodules in the reference reading. Thus, 20 CT scans were used for the analysis. The study was performed under the auspices of a protocol approved by our institutional review board, and the requirement for informed consent was waived.

All CT scans were acquired without an intravenous contrast agent, from the lung apices through the upper abdomen, by using a four–detector row CT scanner (Volume Zoom; Siemens Medical Systems, Erlangen, Germany). Scans were acquired by using a detector configuration of four rows with 1-mm section thickness (4×1 mm), beam pitch of 1.5–1.75, gantry rotation time of 0.5 second, tube potential of 120 kVp, and tube current of 200–300 mA. The data were reconstructed into 1.25-mm-thick sections with 0.6-mm intervals by using a high-resolution reconstruction kernel. The number of sections reconstructed per patient ranged from 431 to 664 (mean, 540).

Image Interpretation by Three Independent Radiologists

Three faculty radiologists (L.C.C., R.M., and A.N.L., with 5, 20, and 10 years of experience, respectively) independently read the 20 CT scans. L.C.C. and R.M. are specialists in general body imaging, and A.N.L. specializes in thoracic imaging. Readings of the transverse CT sections were performed at a standard clinical CT viewing station (Centricity; GE Medical Systems, Milwaukee, Wis) in stacked cine mode. Images were initially displayed with a window level of −750 HU and a window width of 1500 HU, but the readers were free to alter these values at their discretion. The readers were instructed to identify all noncalcified pulmonary nodules with a diameter of 3 mm or more on the CT scans by using a procedure similar to that used in routine clinical practice. The readers used an on-screen cursor that they placed over a nodule candidate to identify its unique three-dimensional coordinates. These coordinates, along with a confidence rating on a scale from 1 to 5 for each nodule candidate, were dictated into a tape recorder. The confidence ratings were as follows: 5, definitely a nodule; 4, probably a nodule; 3, possibly a nodule; 2, unlikely to be a nodule; and 1, very unlikely to be a nodule. The readers timed the interpretation of each patient study with a stopwatch. The recorded data were transcribed onto a spreadsheet by using software (Excel version X for Macintosh; Microsoft, Redmond, Wash) for analysis.

Detailed descriptions of the CAD method (Fig 1), the lung nodule evaluation platform used in establishing the reference standard, the procedure for establishing the reference standard, and the assessment of CAD algorithm performance with leave-one-out cross valida-

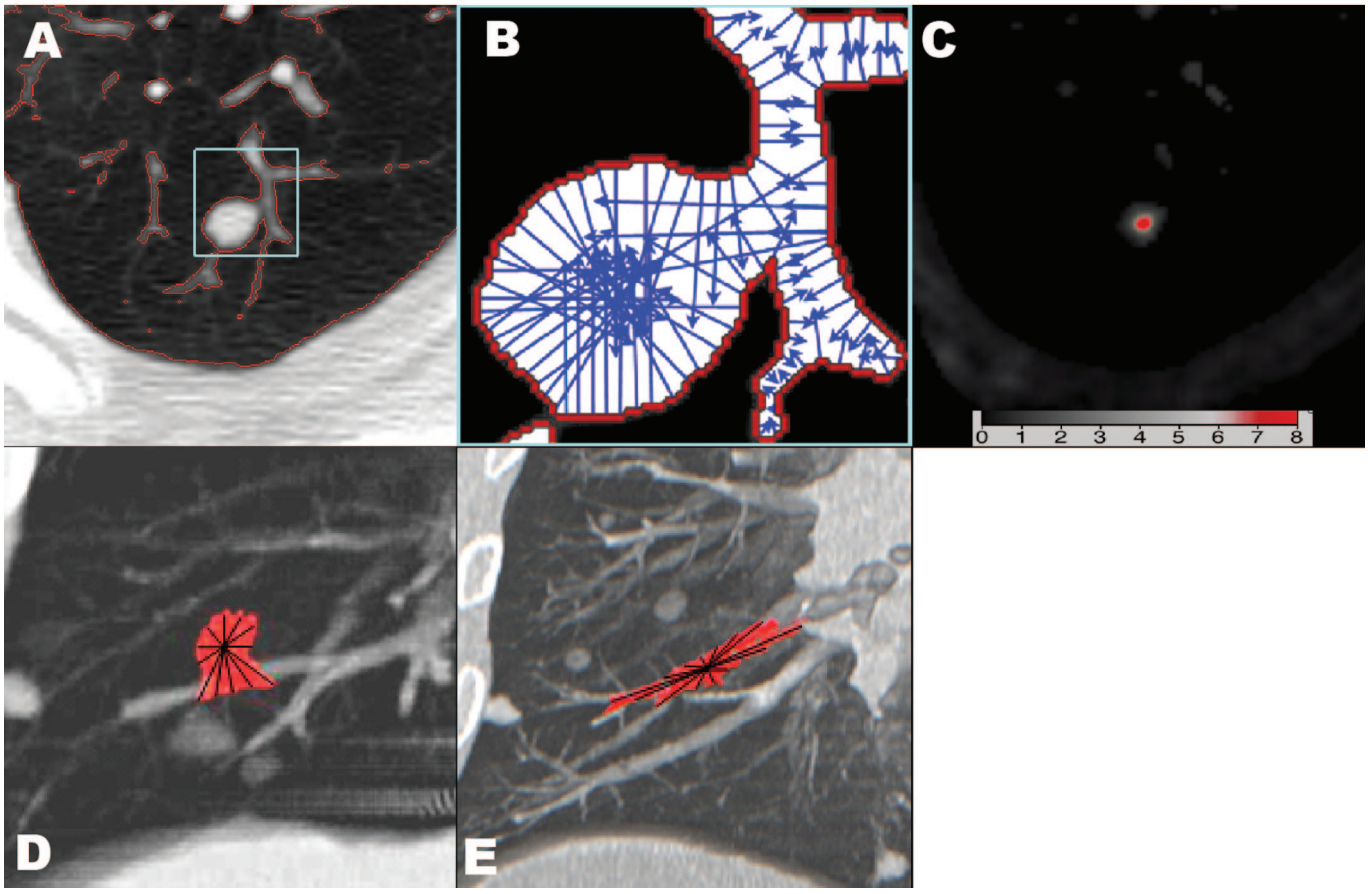

Figure 1. Application of surface normal overlap (SNO)–CAD and lantern transform to nodule detection in human lung. *A,* Detail of transverse CT section shows pulmonary nodule in posterior portion of right upper lobe. Applied threshold indicates isosurface (red) between air and tissue. *B,* Schematic representation of surface normal vectors (blue) generated at isosurface (red) in same lung region as that bounded by aqua box in *A*. Normals are generated in three dimensions in volumetric CT data at all isosurfaces in the thick region extracted during segmentation. *C,* Map of SNO-CAD scores for all voxels displayed in *A,* calculated by using a clustering algorithm to quantify convergence of normal vectors, confirms identification of pulmonary nodule. Scale at bottom indicates 12-bit scaling of SNO-CAD scores (0–8). *D,* Image with superimposed schema shows application of lantern transform to nodule in contact with blood vessel. Rays of visibility (black radii) cast from a SNO-CAD–identified nodule candidate are used to generate an approximately spherical surface (red). *E,* Image with superimposed schema shows application of lantern transform to SNO-CAD–identified nodule candidate in pulmonary vessel. Rays of visibility generate an ellipsoid surface less spherical than that in *D*. Quantitative characteristics of this ellipsoid result in its rejection as a nodule candidate, while those of the ellipsoid in *D* allow it to remain a nodule candidate.

tion are presented in an online-only Appendix to this article (Appendix E1, *radiology .rsnajnls.org/cgi/content/full/2341040589 /DC1*).

Reference Standard

A consensus panel of two thoracic radiologists (D.P.N., G.D.R., with 25 and 14 years of experience, respectively, in reading thoracic CT scans) interpreted the 20 CT studies by using a specially developed computer-based lung nodule evaluation platform, which is described in Appendix E1. The interpretation of the CT scans for establishment of the reference standard occurred both as a free search through the CT sections and as a directed analysis of all nodule candidates identified by the three independent radiologist readers and by the CAD system (training of the CAD system for this purpose is described in Appendix E1). The panel members, without knowledge of the source of detection of the nodule candidates, assessed each candidate and arrived at a final decision in consensus as to whether it was a nodule (TP finding), FP finding, or indeterminate. All FP detections were classified in one of five categories, as peripheral vessel, central vessel, airway wall, amorphous parenchymal opacity, or artifact. In addition, the greatest dimension of each nodule (TP detection) was measured with digital calipers.

Statistical Analyses

The results of CAD after cross validation, and the results of the three individual readings, were compared with the reference standard. Free-response ROC curves were calculated from these data to enable comparison of the performance of radiologists and of CAD by means of alternative free-response ROC analysis (9–11). Although free-response ROC and alternative free-response ROC analyses are not yet as fully developed as classic ROC analysis and are based on certain statistical assumptions (12), they are better suited to the free-response paradigm that was used in this study (9,11,13), and they are widely used in the evaluation of CAD (14). For the three radiologists, free-response ROC curves were created by plotting sensitivity for TP detections versus the average number of FP detections per patient at each of the five operating

points: 5, 4–5, 3–5, 2–5, and 1–5. For CAD alone, the free-response ROC curve was calculated by varying the lower limit of the SNO-CAD performance score (defined in Appendix E1) and similarly plotting sensitivity for TP detection versus the average number of FP detections per patient across a range of SNO-CAD score thresholds. For the comparison of SNO-CAD scores with reader confidence levels by using alternative free-response ROC analysis, we mapped SNO-CAD scores for both TP and FP detections to radiologists' mean performance values (mean FP detections, mean assessment time) at the five confidence levels (First Mapping, Table 1). For example, at the highest confidence level, the mean number of FP detections per patient by the three radiologists was 0.43. Thus, all TP detections made by the CAD system at a threshold above a mean of 0.43 FP detections per patient were mapped to a confidence score of 5. Software (ROCKIT; C. Metz, University of Chicago, Chicago, Ill) was used to analyze alternative free-response ROC data. This software was used to fit a binormal alternative free-response ROC curve to the data from each reader (including the scaled SNO-CAD scores) and to compare the areas under the curve for each pair by using a univariate *z*-score test (15). The procedure for coding free-response ROC data for use with the binormal model used in this software has been described previously (9–11).

To quantify radiologists' potential for achieving increased sensitivity with CAD as opposed to a second radiologist, we compared weighted κ values for radiologist-radiologist and radiologist-CAD agreement. Our hypothesis was that radiologists and CAD tended to detect different groups of nodules. This hypothesis was tested by comparing the overlap between the sets of nodules detected by CAD and by radiologists with the overlap between the sets of nodules detected by different radiologists. For the calculation of weighted κ values for CAD, SNO-CAD scores were mapped onto the 0–5 confidence scale by using two mapping procedures. In the first mapping, the same thresholds described previously for alternative free-response ROC analysis were used. In this mapping procedure, CAD performance results were scaled to radiologist performance results, as was necessary for a quantitative comparison of CAD and radiologist free-response ROC curves by using alternative free-response ROC analysis. This first mapping, however, was focused on a very narrow range of CAD performance parameters that did not reflect how CAD might be used in clinical practice. In the clinical setting, we believe, the time spent in reviewing CAD results will dictate the use of CAD. In our experience, the amount of time that radiologists spend in using the lung nodule evaluation platform to interact with CAD before classifying a CAD-identified nodule candidate as a TP or FP detection is 5–7 seconds. The range of total CAD interaction times in the first mapping was 3–15 seconds. Most radiologists at our institution are willing to spend 1–2 minutes in assessing CAD detections after completing their initial independent review of a CT scan. Therefore, we performed a second mapping that was based on the amount of time that radiologists likely would be willing to spend in assessing CAD detections after their initial independent reading. In this mapping, ratings of 5, 4, 3, and 2 were assigned to interaction times of 30, 60, 90, and 120 seconds, respectively. A rating of 1 corresponded to total interaction times of more than 2 minutes for assessment of all CAD detections. If we assume an average interaction time of 6 seconds per CAD detection, this translates into a mapping based on ranges of 1–5, 6–10, 11–15, 16–20, and more than 20 CAD detections (Second Mapping, Table 1).

Finally, to assess the potential improvement in radiologists' sensitivity with use of CAD as a second reader, we compared the sensitivities (based on the numbers of TP detections at confidence levels of 3–5) of individual radiologists' interpretations to those of double reading and paired radiologist-CAD readings at several CAD thresholds for FP detection. For this analysis, we assumed that radiologists would accept all TP detections made with CAD. Sensitivity levels achieved by radiologists with individual reading, double reading, and CAD were compared by using the McNemar test. For all statistical testing results, a *P* value of less than .05 was considered to indicate a significant difference.

TABLE 1
Mapping of CAD Detections to Reader Confidence Ratings

Confidence Rating	First Mapping		Second Mapping	
	Reader Mean FP Detections per Patient*	CAD FP Detections per Patient	Assessment Time (sec)†	CAD Detections per Patient
5	0.43	≤0.43	≤30	0–5
4	0.80	0.44–0.80	31–60	6–10
3	1.42	0.81–1.42	61–90	11–15
2	1.77	1.43–1.77	91–120	16–20
1	1.82	1.78–1.82	>120	>20
0	. . .	>1.82	. . .	. . .

* Average numbers of FP detections for the three readers.
† Range of times required for radiologist assessment of CAD results.

RESULTS

Reference Standard

The consensus panel was presented with 1297 detections for directed review from the three radiologists and CAD. Of these 1297 detections, the consensus panel characterized 936 as definitely not nodules, 34 as indeterminate, and the remaining 327 as nodules. Of these 327 nodules, 288 were noncalcified nodules with a diameter of 3 mm or more (n = 193) or a diameter of less than 3 mm (n = 95), and 39 were calcified nodules with a diameter of less than 3 mm (n = 17), 3–6 mm (n = 21), or 10 mm (n = 1). In addition to the 193 noncalcified nodules with a diameter of 3 mm or more that were present in the set of radiologist-CAD detections, two other noncalcified nodules within the same diameter range were detected with a free search by the consensus panel, which resulted in an increase in the total number of noncalcified nodules with a diameter of 3 mm or more, to 195. Indeterminate findings, calcified nodules, and noncalcified nodules with a diameter of less than 3 mm represented 13% of all detections and were not included in the performance analysis; in other words, if the reader or the CAD system detected a nodule candidate that met one of these criteria, the observation was not considered in calculations of sensitivity or FP metrics.

We determined the distribution of the 195 noncalcified nodules with a diameter of 3 mm or more across the 20 subjects (Fig 2). The number of nodules per CT scan ranged from 0 to 65. We calculated a histogram of nodule diameters for all 195 nodules identified in the reference read-

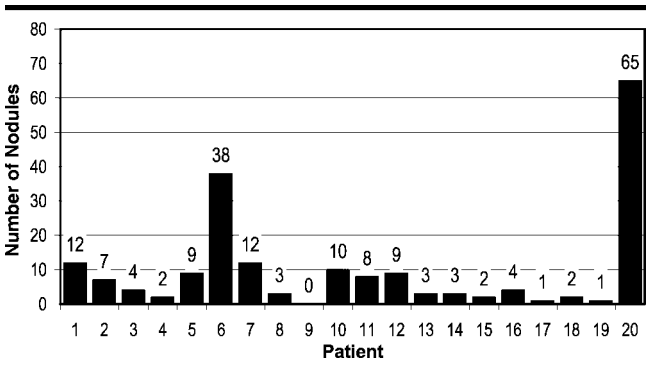

Figure 2. Bar graph shows the per-patient distribution of 195 noncalcified nodules with a diameter of 3 mm or more.

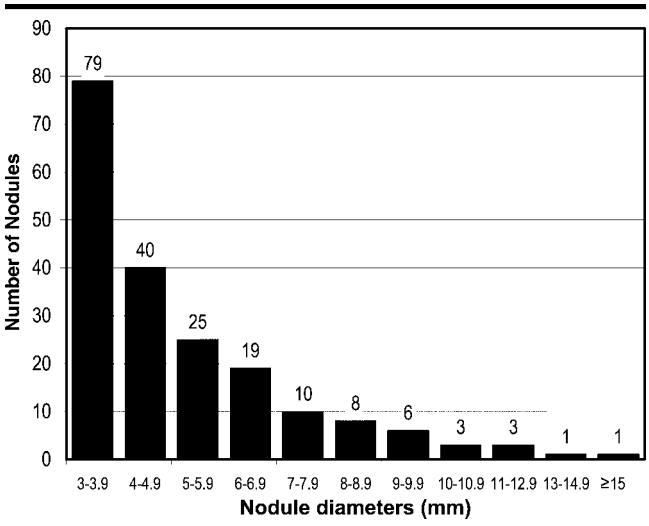

Figure 3. Bar graph shows the distribution of 195 noncalcified nodules with a diameter of 3 mm or more, according to diameter range.

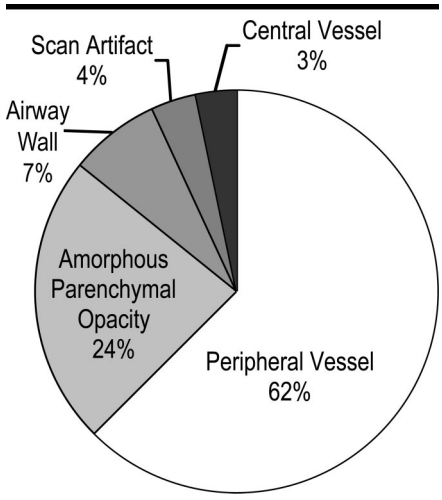

Figure 4. Pie chart shows the distribution of 936 FP detections, determined by the consensus panel to be definitely not nodules, according to the nature of the finding.

ing (Fig 3). The mean nodule diameter was 5.1 mm ± 2.3 (standard deviation). Seventy-one percent of the nodules were located in the peripheral two-thirds of the lung and 29% were located in the central one-third of the lung.

The majority of FP detections (n = 584) were small peripheral vessels, most commonly branch points (Fig 4). Amorphous parenchymal opacities representing platelike atelectasis, scar, or other parenchymal heterogeneity accounted for 220 FP detections, 124 (56%) of which were found in just four patients. Two of these patients had moderate centrilobular emphysema, and the other two had scar or atelectasis in multiple regions of the lung. The remaining FP detections consisted of local airway wall thickening, typically at bifurcations (n = 67); scanning artifact due to respiratory or cardiac motion-induced misregistration (n = 35); and central vessels (n = 30). In general, FP detections were easily discriminated from TP detections by the consensus panel, particularly when the lung nodule evaluation platform was used to observe the three-dimensional relationships of the blood vessels and airways.

Radiologist Performance

Free-response ROC curves of the three radiologists' individual performance were plotted for the 195 noncalcified nodules with a diameter of 3 mm or more and the 76 nodules with a diameter of 5 mm or more (Fig 5). Readers 1–3 required a mean of 4.7, 9.8, and 7.4 minutes (range, 1.8–20.0, 5.0–21.0, and 4.9–20.0 minutes), respectively, to interpret each scan. Reader 2 had the highest sensitivity for detection of nodules with a diameter of 3 mm or more, but spent 108% (5 minutes) and 33% (3 minutes) more time than did readers 1 and 3, respectively, in interpreting each case. It is noteworthy that reader 3 had substantially lower sensitivity, in spite of using 56% more time than did reader 1 to interpret the scans. While striking, this variability is consistent with that in previously published reports of lung nodule detection in animal models (16).

The area under the alternative free-response ROC curve (A_1) was 0.48, 0.55, and 0.36 for readers 1, 2, and 3, respectively (Fig 6). It is important to note that an area under the curve of 0.5 in alternative free-response ROC analysis does not have any special meaning, unlike traditional ROC analysis, in which an area under the curve of 0.5 indicates a performance difference attributable to chance (10). There was a significant difference in area under the alternative free-response ROC curve between reader 3 and the other two readers ($P < .05$), whereas the difference between readers 1 and 2 was not significant ($P = .06$).

Performance of CAD versus Radiologists

The free-response ROC plots for CAD of nodules with a diameter of 3 mm or more and those with a diameter of 5 mm or more are also presented in Figure 5 over the range of per-patient FP detections, with an upper limit of 100 FP detections. The performance of CAD approximates that of the three radiologists over the range of FP detections made by radiologists. Alternative free-response ROC analysis of CAD performance resulted in an area under the curve of 0.54 (Fig 6). Note that this curve represents CAD performance up to a threshold of only 1.82 FP detections per CT scan and not the full

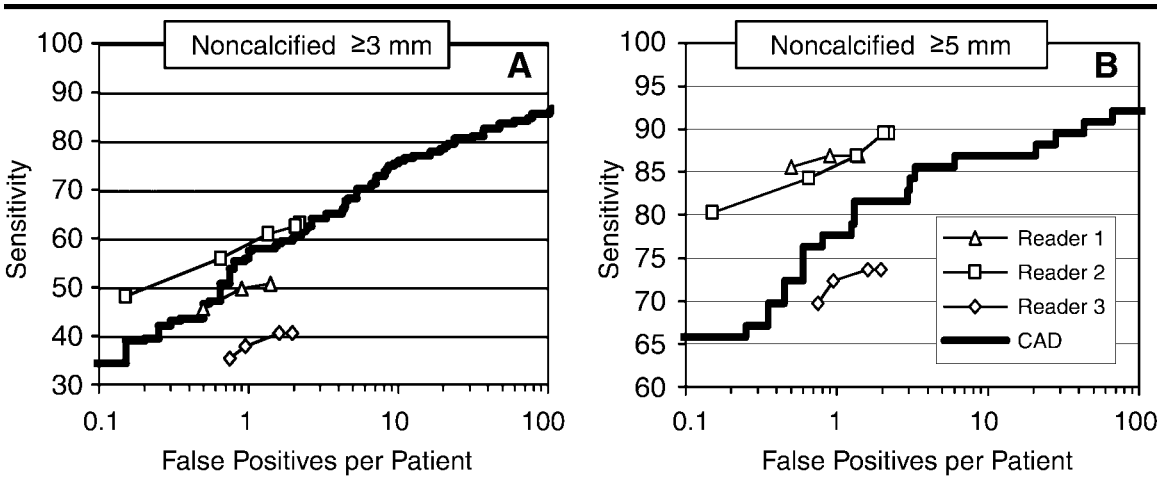

Figure 5. Free-response ROC plots show sensitivity versus FP detections per patient (log scale) for radiologists and the cross-validation–trained CAD system among, *A,* noncalcified nodules with a diameter of 3 mm or more and, *B,* noncalcified nodules with a diameter of 5 mm or more. The legend in *B* applies also to *A,* but note the difference between *A* and *B* in the scaling of the y-axis.

range of CAD detections represented in free-response ROC plots (Fig 5). Alternative free-response ROC analyses in which the performance of each reader was compared with that of CAD indicated that the area under the curve for CAD performance was significantly better than that for reader 3 ($P < .05$). The performance of the CAD system was not significantly different from that of the other two radiologists ($P = .27$ and .78 for readers 1 and 2, respectively).

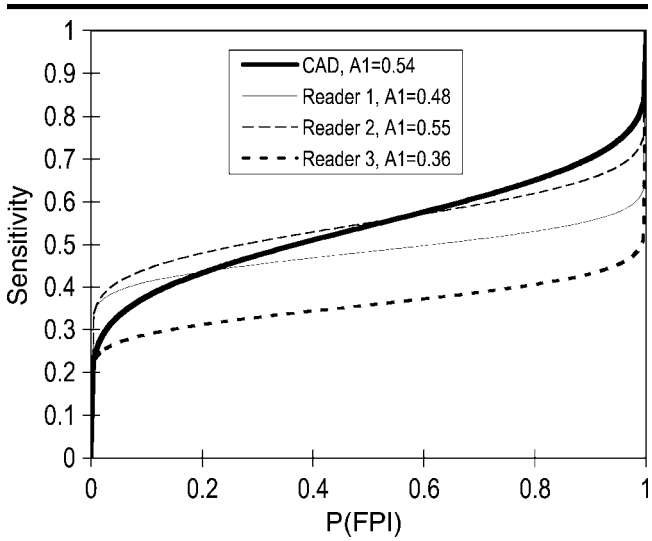

Figure 6. Alternative free-response ROC curves for each reader and CAD. Areas under the curves (*A1*) are given. *P(FPI)* = probability of a FP identification.

Agreement between Reader-Reader and Reader-CAD Interpretations

We prepared two Venn diagrams to illustrate the substantial interreader variability in detection of nodules with a diameter of 3 mm or more and those with a diameter of 5 mm or more at a confidence level of 3–5 for radiologist detections and at an average of 15 FP detections for CAD (Fig 7). This reader confidence level range was selected because reader confidence levels of 3 and higher correspond to lesions that are considered clinically reportable. The CAD threshold was selected to correspond to a hypothetical reader-CAD interaction time of 90 seconds, based on an assumed average of 6 seconds per CAD detection. CAD identified 35 nodules with a diameter of 3 mm or more that were not detected by any of the three readers. CAD also identified the overwhelming majority of nodules detected by the readers. Although the overall performance level of reader 3 was significantly below those of the other two radiologists in this study (Fig 6), it is consistent with other published reports of radiologists' performance (16). In addition, reader 3 identified four nodules with a diameter of 3 mm or more that were not detected by the other readers in our study.

To quantify the differences in performance between readers and CAD, we compared weighted κ values for the three reader-reader combinations and the three reader-CAD combinations by using two different scales for mapping CAD results to reader confidence levels of 0–5 (Table 2).

All three κ values associated with interreader agreement are higher than all three κ values associated with reader-CAD agreement with either of the two mappings. Despite having only three reader-reader and three reader-CAD comparisons, this pattern indicates a significant difference in agreement (Wilcoxon rank sum test, $P < .05$).

Of particular note is that even with the use of strict thresholds for the first CAD mapping, which included only CAD detections associated with fewer than 1.82 FP detections by CAD per case, the CAD system found a mean of 42 lesions (or 2.1 lesions per case) that were missed by radiologists, whereas individual radiologists found a mean of 23 lesions (or 1.15 lesions per case) that were missed by the other reader.

When this interaction was modeled, a

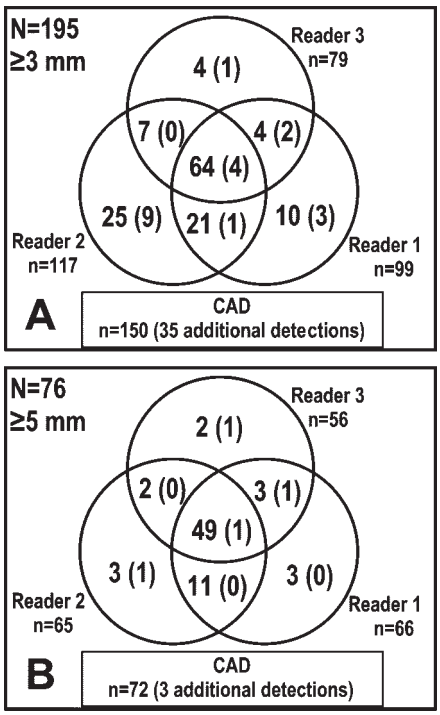

Figure 7. Venn diagrams demonstrate variability among the three readers in the number of TP detections with a confidence level of 3–5 for, *A,* nodules with a diameter of 3 mm or more and, *B,* nodules with a diameter of 5 mm or more. Inside the circles, numbers in parentheses represent TP detections that were not detected by the CAD system. Inside the boxes, numbers indicate total TP detections with CAD used at an upper threshold of 15 FP detections per patient; numbers in parentheses indicate TP detections by the CAD system that were not seen by any reader. These diagrams show that at the selected thresholds all three readers made unique detections and that only 33% of nodules with a diameter of 3 mm or more and 64% of nodules with a diameter of 5 mm or more were detected by all three readers. The CAD system detected a majority of the readers' detections and made 35 unique detections (≥3-mm-diameter nodules)—more than any reader.

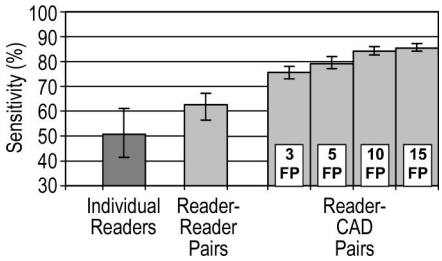

Figure 8. Bar graph shows mean sensitivities for the three readers individually, paired in double readings, and paired with the CAD system, assuming 100% acceptance of TP CAD detections by the radiologists. Four CAD thresholds were tested and are indexed according to the mean number of FP detections that readers would need to assess and exclude. Only nodules identified with a reader confidence level of 3–5 were included in this analysis. The error bars indicate the low-high range of the values. This diagram illustrates that when CAD operates at a threshold that results in only three FP detections per patient, significantly better performance and reduced interradiologist variability (narrower high-low error bars) were found for radiologists paired with the CAD system, compared with radiologists reading alone or in pairs. Note the small incremental improvement in performance as CAD is allowed to become more sensitive at the expense of a greater number of FP detections.

TABLE 2
Comparison of Reader–Reader and Reader–CAD Interpretations

A: Reader–Reader Interpretations

Reader Pair	Weighted κ Value
1 and 2	0.574
1 and 3	0.567
2 and 3	0.486

B: Reader–CAD Interpretations

Reader	Weighted κ Value for First Mapping	Weighted κ Value for Second Mapping
1	0.382	0.190
2	0.360	0.225
3	0.319	0.103

mean sensitivity increase was seen for the detection of noncalcified nodules with a diameter of 3 mm or more. Modeling was performed with pairing of the results of individual radiologists' readings (confidence levels 3–5) by using the "OR rule" and with pairing of the results of individual radiologists' readings with CAD results at various score thresholds (Fig 8). Double reading (reader-reader interpretation) at these confidence levels resulted in a mean of 2.8–3.0 FP detections per patient, depending on the reader pair. The mean sensitivity of individual readers' interpretations was 50% (range, 41%–60%). On average, double reading improved this value to 63% (range, 56%–67%). If the readers accepted all TP CAD detections, the mean sensitivity for reader-CAD interpretations at the different CAD classifier thresholds would be 76% (range, 73%–78%), 79% (range, 77%–82%), 84% (range, 83%–86%), and 85% (range, 84%–87%) for thresholds allowing an average of three, five, 10, and 15 FP detections, respectively. Sensitivity was significantly higher for reader-CAD interpretations than for interpretations by readers individually or in pairs ($P < .05$).

DISCUSSION

Although chest CT is substantially more sensitive for the detection of lung cancer than is chest radiography (17–19), current CT practice (use of 5–10-mm-thick sections for interpretation) has major shortcomings. In the Anti–Lung Cancer Association screening program (18), seven (32%) of 22 lung cancers with diameters ranging from 4 to 13 mm (mean, 8 mm) were initially missed and were diagnosed only in retrospect, after they had grown, at follow-up CT imaging (20). In 281 patients undergoing lung volume reduction surgery for the treatment of advanced emphysema, only eight (47%) of 17 bronchogenic carcinomas were detected at preoperative CT; missed lung cancers had a mean diameter of 8 mm (21). With regard to resection of pulmonary metastases, sensitivity of conventional CT with the use of 8–10-mm-thick sections ranged from 58%–78% relative to a reference standard determined by surgical palpation and excision (22–25). In two recently published series in which the performance of helical CT with 5-mm collimated sections was compared with that of surgical excision and histologic confirmation, the average sensitivity of CT was 70%–75% for all metastases but was substantially lower for smaller lesions (26,27). In an in vivo study of 5-mm-thick CT sections obtained in four dogs with 132 osteosarcoma metastases macroscopically evident at pathologic examination, 10 radiologists had sensitivities from 11%–42% (mean, 30%) in the detection of metastases, of which nine were more than 10 mm, 24 were more than 5–10 mm, and 99 were less than 5 mm in diameter (16). These limitations in sensitivity are due in part to the use of thick CT sections, which limits the detection of smaller pulmonary nodules because of volume averaging. The routine use of thin-section CT (1-mm section thickness), made possible by the development of multi–detector row CT scanners, should substantially improve the detection of lung nodules (28).

As recently as 1998, the most advanced CT scanners were single–detector row CT scanners that required 25–30 seconds to image the entirety of the lungs with 7–10-mm-thick sections. Today, multi–detector row CT with 16 detector rows allows the entire adult lung to be scanned with 1-mm-thick sections in as

little as 5 seconds. This improvement has the potential to increase diagnostic accuracy in pulmonary nodule detection substantially, compared with that at single–detector row helical CT.

Criticisms of the routine acquisition of thin (1-mm) sections for lung imaging have centered around concern for the resultant increase in radiation exposure to the patient, compared with that during the acquisition of thick sections. In the context of multi–detector row CT, this concern is spurious for several reasons. First, 16–detector row CT scanners cannot be used to acquire raw projection data in section thicknesses of more than 1 mm (or 1.5 mm, depending on the manufacturer). The reconstruction of thicker sections is simply the result of the weighted addition of raw projection data acquired with narrow detector widths. Therefore, the radiation exposure associated with a stack of 1-mm sections will be identical to that for the thicker reconstructions. Second, although thin-section reconstructions are noisier than thick ones, it may not be necessary to increase the radiation dose to compensate for these levels of increased noise, as a concomitant reduction in volume averaging improves lesion visibility (29).

Although it might seem evident that the sensitivity of radiologists for the detection of lung nodules would increase with the substantially higher spatial resolution of multi–detector row CT data compared with that of single–detector row CT data, improved sensitivity cannot be assumed, for two main reasons. First, normal lung blood vessels appear more nodule-like in thinner cross-sections, and, thus, differentiation of nodules from vessels is more difficult. Second, a seven- to 10-fold increase in the amount of data to be reviewed per patient could substantially tax the attentiveness required for accurate lung nodule identification. These phenomena may explain why the performance of the three radiologists in the current study was not substantially better than that of radiologists in other studies reported in the literature, in which thicker sections were used for interpretation (16,18,20–27). Thus, current CT technology presents new challenges and opportunities to radiologists that should fundamentally alter the paradigms with which lung scans are acquired and interpreted. We believe that CAD will play a key role in enabling radiologists to maximize their diagnostic performance when interpreting large, thin-section multi–detector row CT scans.

TABLE 3
Comparison of Results of Current Study with Results of Previously Published Studies of CAD on Whole-Lung CT Scans

Study	No. of Nodules	Nodule Diameter (mm)*	Sensitivity (%)	No. of FP Findings per Section	No. of FP Findings per Patient
Armato et al (39)	50	11.5	80	1.0	28
Armato et al (35)	171	6.5	70	1.5	42
Brown et al (33)	36	5–30†	86	NA	11
Gurcan et al (37)	63	8.9	84	5.5	NA
Lee et al (36)	98	5–30†	72	1.1	31
Wormanns et al (38)	68	7.9‡	38	0.1	5.8
Current study§					
Nodules with diameter ≥3 mm	195	5.1	65	0.005	3
			76	0.019	10
			84	0.093	50
Nodules with diameter ≥5 mm	76	7.3	86	0.005	3
			87	0.019	10
			91	0.093	50

Note.—Previously published studies were performed with thick-section single–detector row CT. NA = not applicable.
* Numbers indicate mean unless otherwise specified.
† Numbers indicate range.
‡ Estimated mean from information given in the article.
§ Results of current study (sensitivity and number of FP findings per section and per patient) are given for three points on the free-response ROC curves.

Previous investigations of CAD of nodules on lung CT scans have been reported with sensitivity and FP values for various CAD methods applied to single–detector row CT scans with 5–10-mm-thick sections (30–40) (Table 3). The first study listed in Table 3 is noteworthy for its cohort composed exclusively of patients undergoing low-dose screening CT, in whom lung cancer was present but was either undetected or misinterpreted as benign by a radiologist (39). These data, while obtained in a different patient population and with a different CT technique (10-mm-thick sections) than data in our study, showed that 23 (61%) of 38 undetected lung cancers were missed because the lesions were not seen, while 15 (39%) of 38 undetected cancers were found but were characterized as benign on the basis of their appearance.

Our data demonstrate that it is possible to develop a CAD algorithm that, when applied to thin-section multi–detector row CT scans, has sensitivity comparable with that of radiologists at thresholds that result in only a small number of additional FP detections, which radiologists presumably would be able to eliminate correctly. Moreover, the TP detections made by our CAD system complemented radiologists' TP detections to a greater extent than did those made by second readers and allowed a substantially greater improvement in radiologist sensitivity and reduction in interradiologist variability than did double reading by radiologists.

An important difference between our data and those published previously (Table 3) is that the average diameter of the pulmonary nodules in our cohort was substantially smaller. This is likely attributable to two factors: the thinner-section multi–detector row CT data available for the reference reading, which allowed confident detection of smaller (3–5-mm) lesions, and the use of a reference standard based on observations made by five radiologists and the CAD system. The relevance of the first of these two factors is supported by a previously published study of CAD applied to thin-section single–detector row CT data, which also demonstrated smaller mean nodule diameters on thin sections than on 5–10-mm-thick sections (40). In that study, Brown and colleagues assessed nodule detection on 1-mm-thick single–detector row CT sections in 20-mm-long subvolumes of chest CT data obtained in 29 patients with suspected pulmonary nodules observed initially on thick-section CT scans. The CAD system that they used achieved 100% sensitivity for the detection of 22 nodules with a diameter of 3 mm or more (mean, 6.3 mm) with 15 FP detections per 20-mm-long subvolume (40). While it is impossible to place these results in the context of a whole-lung scan and of other published studies (Table 3), the successful detection of all nodules with a diameter of 3 mm or more, to our knowledge, was not reported prior to that study and may be in part attribut-

able to the use of thin-section acquisition.

The acronym *CAD* has been used to represent both computer-aided detection and computer-aided diagnosis. Initially, these might be considered identical, but there is an important difference when considering pulmonary nodules and lung CT. While CT is currently the most sensitive noninvasive means for detecting pulmonary nodules, the accuracy of differentiation between benign and malignant noncalcified nodules on the basis of a single CT scan is very low (41). Although positron emission tomography, intravenous iodinated contrast medium uptake, and magnetic resonance imaging have been proposed for differentiation of malignant from benign pulmonary nodules, the current clinical standard for diagnosing malignancy in small lesions is to assess nodule growth on serial CT scans until the nodules reach a size threshold at which biopsy or excision is indicated. As a result, we have focused our CAD development on the detection of noncalcified nodules, leaving the decision of nodule risk to the radiologist. This strategy is consistent with the observation that most errors in diagnosis of lung cancer at CT are related to detection failure (20,41–43). As noted by White et al (42), many detection failures are caused by the "satisfaction-of-search" effect, in which interesting but unrelated findings divert the radiologist's attention from the overlooked tumor. In a study of lung cancers missed on CT scans, White et al found that six (43%) of 14 misses were attributable to this effect, which they believed was far more common in the interpretation of CT scans than in that of chest radiographs. CAD is not susceptible to the satisfaction-of-search effect.

There were several limitations to our study. First, the effect of CAD on the radiologist's interpretation was not measured directly but was inferred by using the principles of double reading and the assumption that radiologists would accept all TP detections by CAD and reject all FP detections by CAD. As a result of this limitation, we cannot determine the actual effect of CAD on FP results. While preliminary data about radiologists' performance suggest that sensitivity gains can be achieved with use of CAD without an increase in the number of FP detections (8,40), the successful rejection of computer-aided FP detections by radiologists must be proved directly.

A second limitation is the lack of an absolute reference standard. Because histologic findings are rarely available for comparison with lung CT scans and because nodules that warrant follow-up may be transient findings, we relied on the consensus of two experts, who used sophisticated tools for two- and three-dimensional visualization to establish the standard for a nodule that should be detected and followed up according to current standards of care. All nodules with a diameter of less than 3 mm were ignored because the confident detection of, and necessity of follow-up for, such small lesions is controversial. We do not know whether our results are generalizable to a lung cancer screening population. We studied CAD in the context of the common task of detecting lung nodules on routine lung CT scans, which is made more challenging by the thin sections inherent in multi–detector row CT acquisitions. Finally, our preliminary experience in establishing a reference standard for lung nodules by means of consensus reading taught us that additional radiologists and CAD will detect additional lesions not identified by members of the consensus panel. This result is not surprising in light of the substantial interobserver variability in nodule detection (Fig 7). We elected to include detections made by CAD and the radiologist readers as specific sites for the consensus panel to review, to maximize the probability that the reference standard would contain all nodules with a diameter of 3 mm or more that were detectable with CT. Because the consensus panel was blinded to the source of the detections (ie, CAD or radiologists) and because the training of the CAD system for the purpose of setting the reference standard was independent of the evaluation of the CAD system with cross-validation techniques, any bias favoring CAD performance should have been minimized.

In summary, we have demonstrated that our CAD algorithm can detect a set of pulmonary nodules complementary to that detected by radiologists and, thus, may allow radiologists to improve the sensitivity of multi–detector row CT for lung nodule detection beyond that achieved with double reading by two radiologists. The complementarity of CAD and radiologist readings supports the view that CAD algorithms such as ours can assist radiologists in the detection of pulmonary nodules but cannot replace them. Future studies must focus on the effect of CAD on radiologists' efficiency and must develop an understanding of what number of FP detections would be acceptable with CAD in clinical practice.

References

1. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241–244.
2. Thurfjell E. Mammography screening: one versus two views and independent double reading. Acta Radiol 1994; 35:345–350.
3. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. Acad Radiol 1996; 3:891–897.
4. Quekel LG, Goei R, Kessels AG, van Engelshoven JM. Detection of lung cancer on the chest radiograph: impact of previous films, clinical information, double reading, and dual reading. J Clin Epidemiol 2001; 54:1146–1150.
5. Anderson ED, Muir BB, Walsh JS, Kirkpatrick AE. The efficacy of double reading mammograms in breast screening. Clin Radiol 1994; 49:248–251.
6. Warren RM, Duffy SW. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. Br J Radiol 1995; 68:958–962.
7. Zheng B, Ganott MA, Britton CA, et al. Soft-copy mammographic readings with different computer-assisted detection cuing environments: preliminary findings. Radiology 2001; 221:633–640.
8. MacMahon H. Improvement in detection of pulmonary nodules: digital image processing and computer-aided diagnosis. RadioGraphics 2000; 20:1169–1177.
9. Chakraborty DP. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. Med Phys 1989; 16:561–568.
10. Chakraborty D. Statistical power in observer-performance studies: comparison of the receiver operating characteristic and free-response methods in tasks involving localization. Acad Radiol 2002; 9:147–156.
11. Chakraborty DP, Winter LH. Free-response methodology: alternate analysis and a new observer-performance experiment. Radiology 1990; 174:873–881.
12. Metz CE. Evaluation of CAD methods. In: Doi K, MacMahon H, Giger ML, Hoffman KR, eds. Computer-aided diagnosis in medical imaging. Amsterdam, the Netherlands: Elsevier Science, 1999; 543–554.
13. Edwards DC, Kupinski MA, Metz CE, Nishikawa RM. Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. Med Phys 2002; 29:2861–2870.
14. Doi K, MacMahon H, Katsuragawa S, Nishikawa RM, Jiang Y. Computer-aided diagnosis in radiology: potential and pitfalls. Eur J Radiol 1999; 31:97–109.
15. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234–245.
16. Waters DJ, Coakley FV, Cohen MD, et al. The detection of pulmonary metastases by helical CT: a clinicopathologic study in dogs. J Comput Assist Tomogr 1998; 22:235–240.
17. Latief KH, White CS, Protopapas Z, Attar S, Krasna MJ. Search for a primary lung neoplasm in patients with brain metasta-

sis: is the chest radiograph sufficient? AJR Am J Roentgenol 1997; 168:1339–1344.
18. Kaneko M, Eguchi K, Ohmatsu H, et al. Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography. Radiology 1996; 201: 798–802.
19. Henschke CI, McCauley DI, Yankelevitz DF, et al. Early lung cancer action project: overall design and findings from baseline screening. Lancet 1999; 354:99–105.
20. Kakinuma R, Ohmatsu H, Kaneko M, et al. Detection failures in spiral CT screening for lung cancer: analysis of CT findings. Radiology 1999; 212:61–66.
21. Hazelrigg SR, Boley TM, Weber D, Magee MJ, Naunheim KS. Incidence of lung nodules found in patients undergoing lung volume reduction. Ann Thorac Surg 1997; 64:303–306.
22. Chang AE, Schaner EG, Conkle DM, Flye MW, Doppman JL, Rosenberg SA. Evaluation of computed tomography in the detection of pulmonary metastases: a prospective study. Cancer 1979; 43:913–916.
23. Friedmann G, Bohndorf K, Kruger J. Radiology of pulmonary metastases: comparison of imaging techniques with operative findings. Thorac Cardiovasc Surg 1986; 34(special issue 2):120–124.
24. McCormack PM, Ginsberg KB, Bains MS, et al. Accuracy of lung imaging in metastases with implications for the role of thoracoscopy. Ann Thorac Surg 1993; 56: 863–865; discussion 865–866.
25. Peuchot M, Libshitz HI. Pulmonary metastatic disease: radiologic-surgical correlation. Radiology 1987; 164:719–722.
26. Ambrogi V, Paci M, Pompeo E, Mineo TC. Transxiphoid video-assisted pulmonary metastasectomy: relevance of helical computed tomography occult lesions. Ann Thorac Surg 2000; 70:1847–1852.
27. Diederich S, Semik M, Lentschig MG, et al. Helical CT of pulmonary nodules in patients with extrathoracic malignancy: CT-surgical correlation. AJR Am J Roentgenol 1999; 172:353–360.
28. Fischbach F, Knollmann F, Griesshaber V, Freund T, Akkol E, Felix R. Detection of pulmonary nodules by multislice computed tomography: improved detection rate with reduced slice thickness. Eur Radiol 2003; 13:2378–2383.
29. Mayo JR, Hartman TE, Lee KS, Primack SL, Vedal S, Muller NL. CT of the chest: minimal tube current required for good image quality with the least radiation dose. AJR Am J Roentgenol 1995; 164: 603–607.
30. Giger ML, Bae KT, MacMahon H. Computerized detection of pulmonary nodules in computed tomography images. Invest Radiol 1994; 29:459–465.
31. Kanazawa K, Kawata Y, Niki N, et al. Computer-aided diagnosis for pulmonary nodules based on helical CT images. Comput Med Imaging Graph 1998; 22: 157–167.
32. Armato SG 3rd, Giger ML, Moran CJ, Blackburn JT, Doi K, MacMahon H. Computerized detection of pulmonary nodules on CT scans. RadioGraphics 1999; 19:1303–1311.
33. Brown MS, McNitt-Gray MF, Goldin JG, Suh RD, Sayre JW, Aberle DR. Patient-specific models for lung nodule detection and surveillance in CT images. IEEE Trans Med Imaging 2001; 20:1242–1250.
34. Ko JP, Betke M. Chest CT: automated nodule detection and assessment of change over time—preliminary experience. Radiology 2001; 218:267–273.
35. Armato SG 3rd, Giger ML, MacMahon H. Automated detection of lung nodules in CT scans: preliminary results. Med Phys 2001; 28:1552–1561.
36. Lee Y, Hara T, Fujita H, Itoh S, Ishigaki T. Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique. IEEE Trans Med Imaging 2001; 20:595–604.
37. Gurcan MN, Sahiner B, Petrick N, et al. Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system. Med Phys 2002; 29:2552–2558.
38. Wormanns D, Fiebich M, Saidi M, Diederich S, Heindel W. Automatic detection of pulmonary nodules at spiral CT: clinical application of a computer-aided diagnosis system. Eur Radiol 2002; 12:1052–1057.
39. Armato SG 3rd, Li F, Giger ML, MacMahon H, Sone S, Doi K. Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program. Radiology 2002; 225: 685–692.
40. Brown MS, Goldin JG, Suh RD, McNitt-Gray MF, Sayre JW, Aberle DR. Lung micronodules: automated method for detection at thin-section CT—initial experience. Radiology 2003; 226:256–262.
41. Li F, Sone S, Abe H, MacMahon H, Armato SG 3rd, Doi K. Lung cancers missed at low-dose helical CT screening in a general population: comparison of clinical, histopathologic, and imaging findings. Radiology 2002; 225:673–683.
42. White CS, Romney BM, Mason AC, Austin JH, Miller BH, Protopapas Z. Primary carcinoma of the lung overlooked at CT: analysis of findings in 14 patients. Radiology 1996; 199:109–115.
43. Gurney JW. Missed lung cancer at CT: imaging findings in nine patients. Radiology 1996; 199:117–122.