flat-panel-detector unit (DigitalDiagnost; Philips, Best, the Netherlands). Images were processed by using nonlinear multifrequency-band processing (13); parameters recommended by the manufacturer were used. For all patients, posteroanterior and lateral projections were available.

The screening CT examinations were performed with 16 × 0.75-mm collimation at 30 mAs and 120–140 kV, depending on weight. Sections of 1 mm thickness were reconstructed every 0.7 mm.

Standard of Reference

In the cancer-positive cases, the exact location of each nodule on a chest radiograph was determined by two observers who did not participate as a reader (B.d.H., radiology researcher with 3 years experience in reading CT lung cancer screening studies). In case of doubt, he consulted an independent chest radiologist (M.P.). This chest radiologist also judged whether lesions were retrospectively visible on a chest radiograph. Both had access to chest radiographs, as well as screening CT scans. Lesions that were, even with the knowledge of the CT findings, not visible on the chest radiograph were excluded from analysis.

Findings of all screening CT examinations were evaluated for nodules according to the criteria set by the lung cancer screening program (14). Volumetric software (Lung Care 5 VB10A-W; Siemens Medical Solutions, Erlangen, Germany) was used to assess nodule volume. This volume was used to calculate the diameter on the basis of the assumption of a perfect sphere.

CAD System

We used a commercially available CAD system (Onguard 5.0; Riverain, Miamisburg, Ohio). The software highlights regions suspicious for containing a focal lung lesion by placing a circle of 5 cm in diameter around the suspicious area (Fig 1). Images are automatically processed in the background so that results are immediately available on demand when the chest radiograph is being read by a radiologist. The program only analyzes the posteroanterior or anteroposterior projection. According to the manufacturer, the algorithm was optimized to detect nodules of 9–30 mm in diameter, although in practice, it also marks larger and smaller nodules.

Figure 1

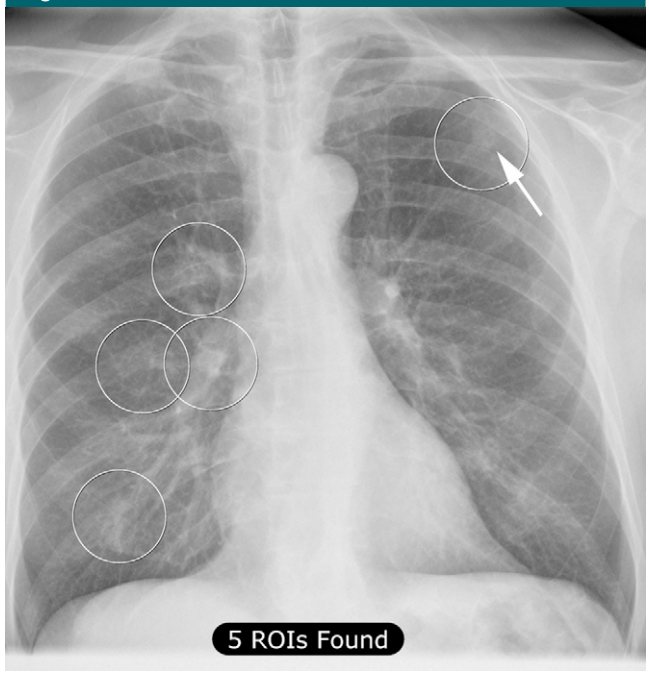


Figure 1: Chest radiograph shows TP (arrow) and FP CAD annotations in a patient with malignancy in the left upper lobe. *ROIs* = regions of interest.

Stand-Alone CAD Performance

To assess stand-alone performance of the CAD system, annotations were labeled TP if the suspicious lesion was located at least partially within the central 50% of the circular CAD annotation.

Observer Study

Images were evaluated on Digital Imaging and Communications in Medicine-calibrated liquid crystal display monitors (MFGD 3220D; Barco, Kortrijk, Belgium) with a matrix size of 2048 × 1536. Options for magnification and adaptation of window settings were available. All chest radiographs were anonymized. Posteroanterior and lateral images were available for evaluation. Chest radiographs were shown in alphabetical order on the basis of patient name to six independent observers. The observers varied in their level of experience: one general radiologist with 6 years of experience (observer A), one chest radiologist with more than 20 years of experience (observer B), and four radiology residents with experience that varied from 1 to 4 years (observers C–F). Observers knew that the study group was chosen from a lung cancer screening trial and they were also told that some patients might have more than one malignant lesion. Two of the observers (reader B and E) had used the CAD system before during other reader studies, but none of the readers had routine experience. To familiarize the observers with the CAD system, five cancer cases that were not included in the observer study were shown to the observers without and with CAD annotations before the start of the study.

Each chest radiograph was first evaluated without and subsequently with CAD results, and observer readings were recorded separately. On a per-patient basis, the observers were asked to document all potentially malignant focal abnormalities seen on the chest radiograph on a separate paper printout with respect to the anatomic lesion locations

Table 1

Demographic Characteristics of Study Participants

Characteristic	Cases ($n = 46$)	Control Subjects ($n = 65$)	*P* Value
Mean age (y)	64.0 (6.0)*	62.5 (5.3)*	.18
No. of men/women	41/5	54/11	.37

*Data in parentheses are standard deviations.

and the readers' confidence scores by using a four-point scale (score of 1: potential lesion, very low degree of suspicion; score of 2: dubious lesion; score of 3: probable lesion; and score of 4, definite lesion). Observers were allowed to mark multiple suspicious lesions on each chest radiograph. They were instructed, however, to ignore nodules smaller than 5 mm in diameter. The researcher (B.d.H.) and the experienced radiologist (M.P.) who had not been involved in the readings analyzed all paper printouts, with the chest radiographs and CT scans being available. The readers' markings were considered TP if the centers of the markings were within the boundaries of the nodules on the chest radiograph. Locations that did not match with a lesion were classified as FP.

Data Analysis

Free-response receiver operating characteristic (FROC) analysis of the observer study was performed as described by Swensson (15) on a per-marker basis. Jackknife FROC, especially developed to analyze observer free-response tasks (16–18), was used to analyze the FROC data. Jackknife FROC software (JAFROC, version 2.3a; *http://www.devchakraborty.com*) (16,19) was used to compute a figure of merit (FOM). The FOM is defined as the probability that lesions (including unmarked lesions) are rated higher than nonlesion marks on control chest radiographs (17), or, in other words, that lesions are given a higher confidence rating for the presence of malignancy than normal findings. Normal images with no marks and unmarked lesions are assigned a zero rating. The level of significance was corrected for multiple comparisons by using Bonferroni correction.

Sensitivity was calculated as the number of TP markings divided by the total number of malignancies. All observer markings, even those that were scored with low confidence, were included to calculate the sensitivity and the FP rate.

Since it is controversial whether application of CAD as a second reader also allows for discharge of candidates seen without CAD (20), we also evaluated a situation in which the observers could only increase their suspicion with CAD while preserving all lesion locations seen without CAD.

In an effort to understand the effect of lesion conspicuity on our results, we performed a separate jackknife FROC analysis on conspicuous nodules, defined as lesions that were detected by three or more readers.

To test for demographic differences between the cases and the control subjects, we compared both groups with respect to sex by using a χ^2 test and age by using a Student *t* test. *P* values less than .05 were considered to indicate a significant difference.

Results

Sample Characteristics

A total of 46 participants with 49 histologically proved pulmonary malignancies met the criteria for the cancer-positive cases. Sixty-five subjects met the criteria for control cases. Indications for acquisition of the chest radiograph in the control group were exclusion of acute cardiovascular disease ($n = 18$), chronic obstructive pulmonary disease ($n = 18$), screening for lung abnormalities because of rheumatoid arthritis ($n = 10$), preoperative screening ($n = 10$), unexplained fever ($n = 4$), chronic cough ($n = 3$), malaise ($n = 1$), and trauma ($n = 1$).

Cases did not differ significantly from the control subjects with respect to age and sex (Table 1). Tumor diameter ranged from 5.1 to 50.7 mm (median, 12.0 mm), with two lesions being larger than 30 mm. Conspicuity of malignancies was very variable: Ten of 49 (20%) malignancies were detected by all six observers without the use of CAD. Furthermore, 11 malignancies were detected by five observers, two were detected by four observers, six were detected by three observers, five were detected by two observers, and seven were detected by only one observer without the use of CAD. Eight (16%) malignancies were not detected by any of the observers without or with the use of CAD. None of the 43 small benign nodules in the control group was marked by either the CAD system or any of the observers.

CAD Stand-Alone Performance

The CAD stand-alone sensitivity was 61% (30 of 49), with on average 2.4 FP annotations (range, zero to five) per chest radiograph. CAD depicted three malignancies that were initially not detected by any of the observers. The diameter of the CAD-depicted malignancies ranged from 7.0 to 50.7 mm.

Observer Performance without CAD

Without CAD, the FOM was 0.72 for radiologists and 0.58 for residents (Table 2, Fig 2). The radiologists had an average sensitivity of 63%, with 0.23 FP annotations per chest radiograph. The residents had an average sensitivity of 49%, with 0.45 FP annotations per chest radiograph. Twenty-seven lesions were detected by at least three observers. In this subselection of more conspicuous lesions, the average FOM was 0.93 for radiologists and 0.76 for residents, with an average sensitivity of 96% for radiologists and 75% for residents.

Observer Performance with CAD When Lowering of Confidence Scores Was Allowed

When the readers were allowed to change their ratings depending on CAD

Figure 2

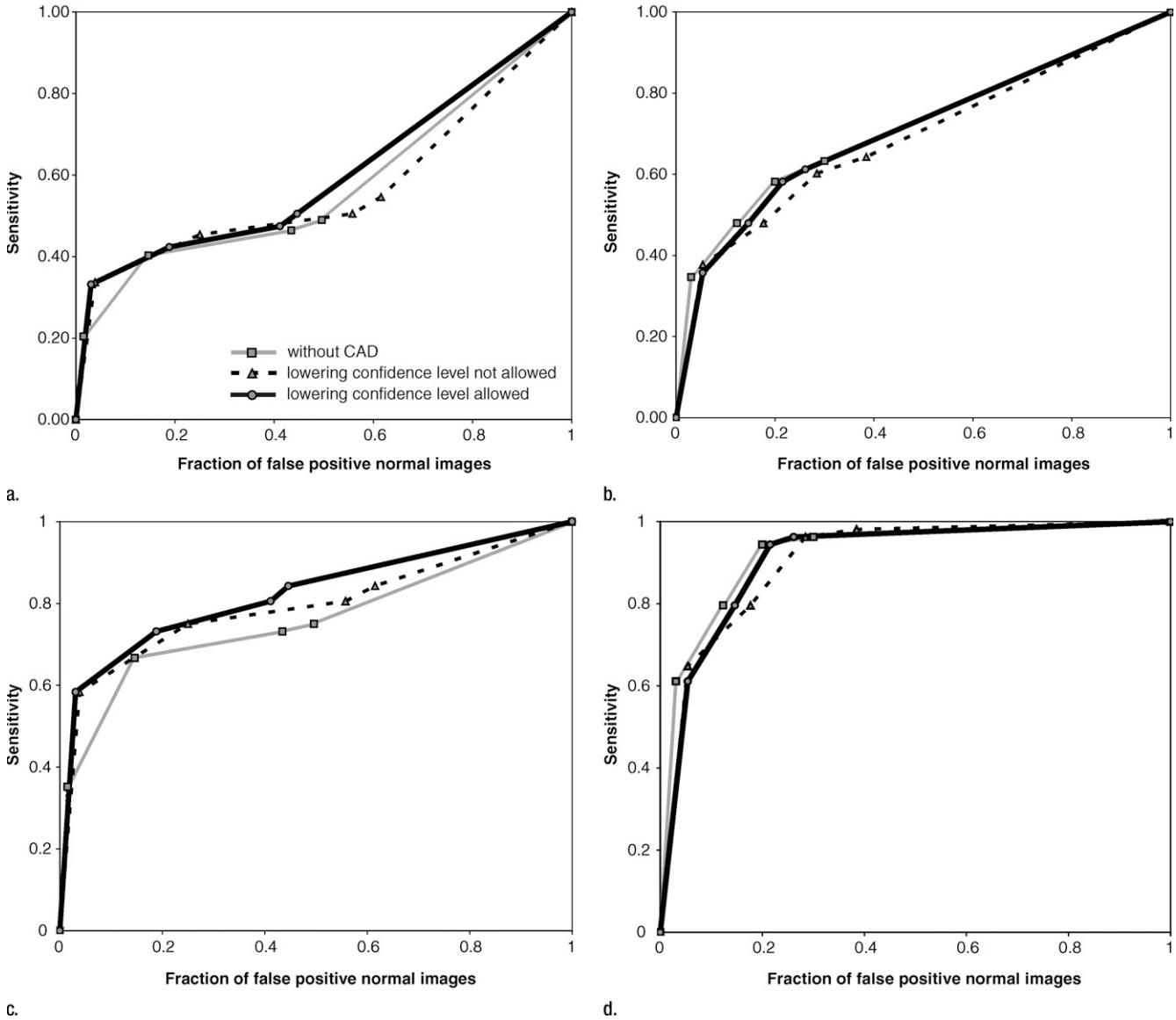


Figure 2: Alternative FROC curves for detection of pulmonary malignancies by **(a, c)** residents and **(b, d)** radiologists. Separate analysis for all lesions (**a, b**) and more conspicuous lesions seen by more than two observers (**c, d**) was performed. The FOM, which is area under the alternative FROC curve, improved significantly for detection of more conspicuous lesions by residents if they were allowed to freely adjust their level of confidence after being provided with the CAD output. The remaining alternative FROC curves did not significantly change with use of CAD.

suggestions, average FOM for the radiologists did not change (0.72, $P = .98$). Average FOM for the residents increased from 0.58 to 0.61, but the improvement was not significant ($P = .08$) (Table 2). With CAD, the average sensitivity of radiologists and residents remained virtually unchanged, 61% and 51%, respectively. Specificity improved, from 0.23 to 0.19 FP annotations per chest radiograph for radiologists and from 0.45 to 0.36 FP annotations for residents.

In the subselection of conspicuous lesions, average FOM remained 0.93 for radiologists, but significantly improved for residents (from 0.76 to 0.82, $P < .001$). Sensitivity remained 96% for radiologists, but improved from 75% to 84% for residents.

Observer Performance with CAD When Lowering of Confidence Scores Was Not Allowed

When readers were only allowed to increase their confidence scores after having CAD results, average FOM decreased from 0.72 to 0.70 for radiologists ($P < .001$) and from 0.58 to 0.57 for residents ($P = .6$). Average sensitivity increased from 63% to 64% (range,

63%–65%) for radiologists and from 49% to 55% (range, 41%–69%) for residents, but the average number of FP annotations per chest radiograph also increased from 0.23 to 0.31 and from 0.45 to 0.54, respectively.

Interaction between CAD and Readers

Together, the six observers placed a total of 66 new markings after having CAD results: 12 for TP CAD annotations and 54 for FP CAD annotations. The number of additionally detected malignancies following TP CAD annotations ranged from zero to six for the various observers (Table 3). The residents benefited more from CAD than did the radiologists, but they also accepted more FP CAD annotations, on average one per 11 chest radiographs versus one per 19 chest radiographs for the radiologists.

Observers A, B, C, D, E, and F, respectively, dismissed 23, 4, 35, 35, 3, and 17 of their own initial markings because CAD had not annotated these regions (Table 4). The number of malignancies initially not seen by the observers but correctly annotated by CAD varied between five and 16 per observer. Eighty percent (47 of 59) of these TP CAD annotations were rejected by the observers (Table 3). An example is shown in Figure 3.

The average confidence levels were generally low for new TP markings, new FP markings, and markings that were initially called but later dismissed after seeing CAD annotations, with confidence levels of 1.9, 1.8, and 1.6, respectively.

Table 2

Individual Outcome of Observer Study without and with CAD When Lowering of Confidence Score Was Allowed

	FOM		Sensitivity (%)		FP Markings per Chest Radiograph	
Reader	Without CAD	With CAD	Without CAD	With CAD	Without CAD	With CAD
Radiologist						
A	0.73	0.75	63	57	0.25	0.11
B	0.71	0.70	63	65	0.22	0.28
Average	0.72	0.72	63	61	0.23	0.19
Resident						
C	0.47	0.53	39	41	0.59	0.41
D	0.60	0.63	69	65	0.75	0.58
E	0.62	0.65	37	41	0.13	0.12
F	0.62	0.62	51	55	0.32	0.34
Average	0.58	0.61	49	51	0.45	0.36

Table 3

Potential of CAD to Improve Observer Performance

	Radiologist			Resident				
Variable	A	B	Average	C	D	E	F	Average
No. of TP CAD annotations initially not detected by observers	5	7	6	13	5	16	13	11.8
No. of rejected TP CAD annotations	5	6	5.5	10	5	14	7	9

Note.—CAD correctly annotated 30 malignancies. Most TP CAD annotations were rejected by the observers.

Table 4

Effect of CAD on the Number of TP, FP, True-Negative, and False-Negative Markings of the Observers

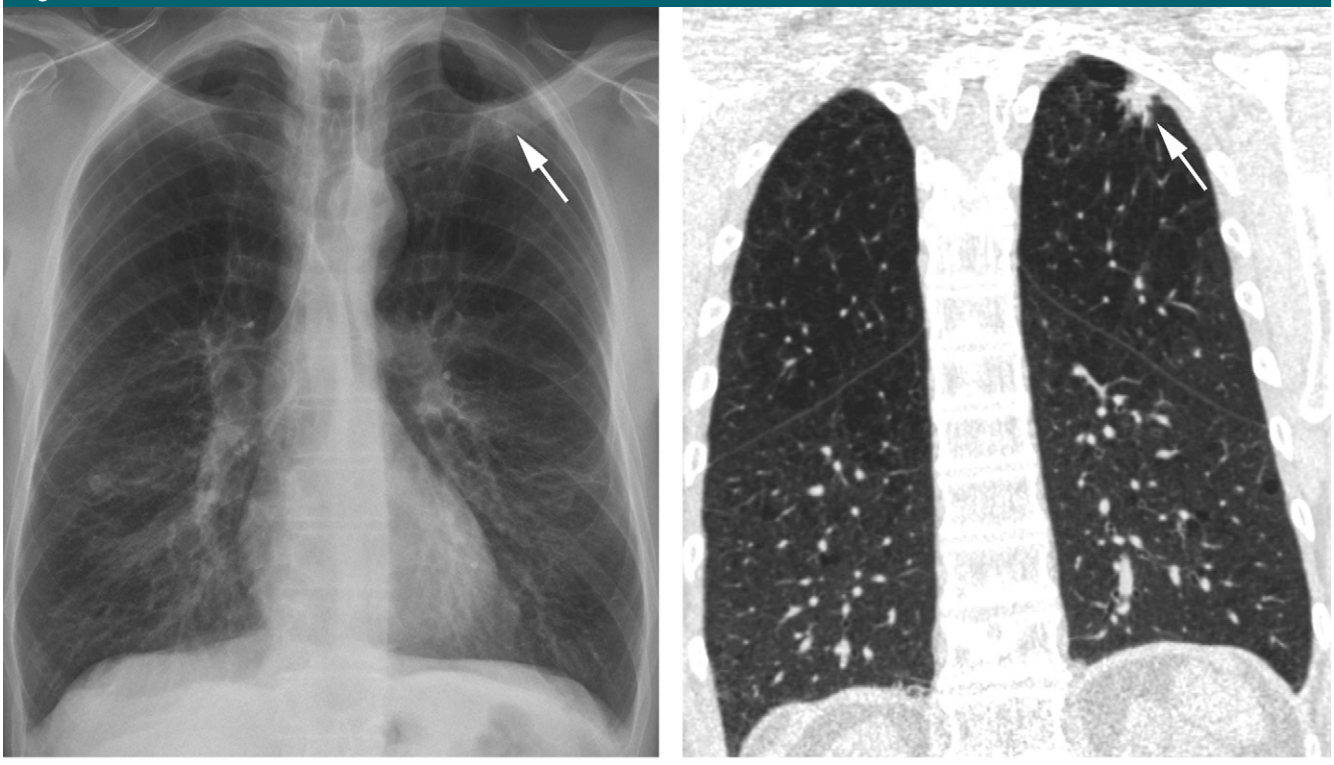
	Radiologist			Resident				
Effect of CAD*	A	B	Average	C	D	E	F	Average
Positive								
FN to TP markings	0	1	0.5	3	0	2	6	2.8
FP to TN markings	20	4	12	33	33	3	13	20.5
Negative								
TN to FP markings	3	9	6	13	13	1	15	10.5
TP to FN markings	3	0	1.5	2	2	0	4	2

*FN = False-negative, TN = true-negative.

Discussion

In this study we assessed how recently released, commercially available CAD software affected reader performance in detecting early lung cancer on chest radiographs. Stand-alone sensitivity of CAD was virtually identical to that of experienced radiologists: 61% in a dataset where 16% of the nodules were not detected by any of the observers. However, the number of FP annotations per chest radiograph was, on average, 10 times higher with CAD than with the two radiologists. The number of CAD-annotated malignancies that were initially not detected by observers varied between five and 16 per observer, out of a total of 49 malignancies, which indicates a vast potential for CAD to improve reader performance. Still, no significant improvement in observer performance could be demonstrated with use of CAD as a second reader in the detection of nodules on chest radiographs.

Figure 3



a. b.

Figure 3: **(a)** Chest radiograph and **(b)** CT scan of correctly CAD-annotated adenocarcinoma (arrow). Both radiologists detected the tumor without CAD, but none of the four residents marked the region, even after seeing CAD results.

An interesting observation is that in the current study, CAD did not improve observer performance. The reason is not that the observers disregarded the CAD annotations; on the contrary, in total, 66 CAD annotations were accepted and 117 initial observer markings were removed because CAD did not annotate the corresponding region. The 66 accepted annotations, pooled over all observers, were 12 TP CAD annotations of lesions initially missed and 54 FP CAD annotations. Among the 117 removed markings were 11 TP lesions. This shows that the observers had difficulties differentiating TP from FP CAD annotations.

This principle has previously been described in a chest radiograph nodule detection study in which eye-tracking was used. In that study, only a minority of the lesions were missed due to inefficient search. The dominant cause of unreported nodules proved to be incorrect decision-making (21). This has also been described in a study that used CAD for detection, as well as classification of suspicious regions (22). The detection function of that CAD system annotated suspicious regions, but only slightly increased the number of lung cancers detected by the observers. Similar to our study, cancers initially missed by the observers but correctly annotated by CAD were frequently rejected by the observers. The authors report that the missed cases were mainly subtle lesions. The reported improvement in radiologists' performance was mainly due to the classification function that computed the likelihood of malignancy for regions indicated by the observer. Using this information, the observer could then change his or her initial decision.

All malignancies included in our study were depicted with CT during lung cancer screening. Malignancies detected during CT screening are usually in an early stage and consequently more difficult to recognize on chest radiographs (23), a fact that is reflected in this study by the relatively low sensitivity of the observers. In a previous study (24) analyzing CAD, pulmonary malignancies that were inadequately visible on chest radiographs were excluded from the analysis. A very high area under the ROC curve of 0.92 was reported without the use of CAD. Observer performance was reported to improve significantly with CAD, and detection became almost flawless. When we excluded subtle lesions from our database and repeated the analysis, we also found excellent performance for radiologists (FOM, 0.93) and significant improvement in FOM to 0.82 for residents. These results show that classification is less problematic in nonsubtle lesions and the benefit of CAD is larger in more conspicuous cases, although such obvious lesions are less likely to be missed in the first place by experienced radiologists.

We showed that to improve observer performance for subtle lesions, observers need to learn to better differentiate between TP and FP CAD annotations. Observer training to recognize FP CAD annotations or a change in how CAD presents results might lead to this goal. In that respect, the lack of training of our observers might have contributed to the low positive effect of CAD. On the other hand, it is, to date, unknown how much training would be necessary and how strong such learning effects would be.

CAD systems of the future may not only provide annotations, but also assign likelihood that an annotation is a true lesion. Alternatively, CAD may just display the likelihood of a suspicious area on demand. This approach has been shown to improve detection of cancers on mammograms (25). However, it requires a low threshold for radiologists to query potentially suspicious regions marked by CAD, and it also will not prevent missing detection errors by the radiologist.

No consensus exists as to whether it is allowed for observers to reduce their suspicion when using CAD. Some state that radiologists should never reduce their initial level of suspicion for markings, irrespective of the CAD results (20). However, the interaction between the radiologist's confidence and the CAD markings is unavoidable in clinical practice, and final diagnosis will be the result of the interaction between the individual reader and CAD. Our separate analysis under the condition that observer ratings could not be reduced after seeing CAD results demonstrated that this approach actually resulted in a 2%–6% higher sensitivity, however, at the cost of such an increase of FP markings that the FOM decreased significantly with this approach.

The relative high number of FP markings in our study can be explained by the low threshold for calling a marking positive. Even the lowest rating was already counted as a positive reading. This threshold is also used in other studies that evaluated CAD (22,26) and ensures that all changes made owing to CAD are accounted for in the evaluation, because observers had a low confidence in most markings that were placed after seeing the CAD results.

Our study was limited by the fact that the observers were explicitly asked to search for lung nodules. In clinical practice, chest radiographs are often requested for other reasons than lung cancer screening. In such cases, the search for lung nodules by radiologists is potentially less thorough, and nodules may be overlooked more easily. CAD may be more beneficial in such a situation. In addition, there was a bias toward calling suspicious abnormalities a lesion, because observers knew that the prevalence of cases was higher than in a normal screening situation. In practice, lower detections rates for the observers may therefore be likely. How far that will affect their attitude toward positive CAD markings is unknown. Finally, although we did not find significant improvement in FOM with the use of CAD, we did find a strong trend for the residents. All residents showed an equal or higher FOM with the use of CAD, with an improvement that reached a P level of .08. It is likely that this improvement would have yielded statistical significance when more cases or observers had been included. More research is needed to confirm this trend for the use of this CAD system by residents.

We conclude that the detection rate of pulmonary malignancies on chest radiographs is comparable for current CAD software and experienced radiologists. However, the positive predictive value of CAD was limited by the high FP rate. Because observers were unable to sufficiently differentiate TP from FP annotations, CAD did not significantly change nodule detection performance. CAD significantly improved detection of more conspicuous lesions by less experienced observers. For subtle lesions, however, additional measures are needed to be able to take advantage of lesions that were missed by observers but were annotated by CAD. Special training of readers might help them differentiate TP from FP CAD annotations. As an alternative, CAD findings might be presented so that they also provide an estimation of the probability of malignancy.

References

1. Spring DB, Tennenhouse DJ. Radiology malpractice lawsuits: California jury verdicts. Radiology 1986;159(3):811–814.
2. Gavelli G, Giampalma E. Sensitivity and specificity of chest X-ray screening for lung cancer: review article. Cancer 2000;89 (11 suppl):2453–2456.
3. Li F, Arimura H, Suzuki K, et al. Computer-aided detection of peripheral lung cancers missed at CT: ROC analyses without and with localization. Radiology 2005;237(2): 684–690.
4. Potchen EJ, Cooper TG, Sierra AE, et al. Measuring performance in chest radiography. Radiology 2000;217(2):456–459.
5. Quekel LG, Kessels AG, Goei R, van Engelshoven JM. Detection of lung cancer on the chest radiograph: a study on observer performance. Eur J Radiol 2001;39(2): 111–116.
6. Toyoda Y, Nakayama T, Kusunoki Y, Iso H, Suzuki T. Sensitivity and specificity of lung cancer screening using chest low-dose computed tomography. Br J Cancer 2008;98(10): 1602–1607.
7. Austin JH, Romney BM, Goldsmith LS. Missed bronchogenic carcinoma: radiographic findings in 27 patients with a potentially resectable lesion evident in retrospect. Radiology 1992;182(1):115–122.
8. Quekel LG, Kessels AG, Goei R, van Engelshoven JM. Miss rate of lung cancer on the chest radiograph in clinical practice. Chest 1999;115(3):720–724.
9. Monnier-Cholley L, Arrivé L, Porcel A, et al. Characteristics of missed lung cancer on chest radiographs: a French experience. Eur Radiol 2001;11(4):597–605.
10. Shah PK, Austin JH, White CS, et al. Missed non-small cell lung cancer: radiographic findings of potentially resectable lesions evident only in retrospect. Radiology 2003;226(1): 235–241.
11. International Early Lung Cancer Action Program Investigators, Henschke CI, Yankelevitz DF, et al. Survival of patients with stage I lung cancer detected on CT screening. N Engl J Med 2006;355(17):1763–1771.
12. van Iersel CA, de Koning HJ, Draisma G, et al. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice

CT screening trial (NELSON). Int J Cancer 2007;120(4):868–874.

13. Stahl M, Aach T, Dippel S. Digital radiography enhancement by nonlinear multiscale processing. Med Phys 2000;27(1):56–65.
14. Xu DM, Gietema H, de Koning H, et al. Nodule management protocol of the NELSON randomised lung cancer screening trial. Lung Cancer 2006;54(2):177–184.
15. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. Med Phys 1996;23(10):1709–1725.
16. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. Med Phys 2004;31(8):2313–2330.
17. Chakraborty DP. Analysis of location specific observer performance data: validated extensions of the jackknife free-response (JAFROC) method. Acad Radiol 2006;13(10):1187–1193.
18. Vikgren J, Zachrisson S, Svalkvist A, et al. Comparison of chest tomosynthesis and chest radiography for detection of pulmonary nodules: human observer study of clinical cases. Radiology 2008;249(3):1034–1041.
19. Zheng B, Chakraborty DP, Rockette HE, Maitz GS, Gur D. A comparison of two data analyses from two observer performance studies using Jackknife ROC and JAFROC. Med Phys 2005;32(4):1031–1034.
20. Giger ML, Chan HP, Boone J. Anniversary paper: history and status of CAD and quantitative image analysis—the role of Medical Physics and AAPM. Med Phys 2008;35(12):5799–5820.
21. Manning DJ, Ethell SC, Donovan T. Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. Br J Radiol 2004;77(915):231–235.
22. Shiraishi J, Abe H, Li F, Engelmann R, MacMahon H, Doi K. Computer-aided diagnosis for the detection and classification of lung cancers on chest radiographs ROC analysis of radiologists' performance. Acad Radiol 2006;13(8):995–1003.
23. Henschke CI; for the International Early Lung Cancer Action Program Investigators. Survival of patients with clinical stage I lung cancer diagnosed by computed tomography screening for lung cancer. Clin Cancer Res 2007;13(17):4949–4950.
24. Kakeda S, Moriya J, Sato H, et al. Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system. AJR Am J Roentgenol 2004;182(2):505–510.
25. Karssemeijer N, Otten JD, Verbeek AL, et al. Computer-aided detection versus independent double reading of masses on mammograms. Radiology 2003;227(1):192–200.
26. Sakai S, Soeda H, Takahashi N, et al. Computer-aided nodule detection on digital chest radiography: validation test on consecutive T1 cases of resectable lung cancer. J Digit Imaging 2006;19(4):376–382.