TABLE 1 Values for Area Under Receiver Operating Characteristic Curve (A_z) of 16 Radiologists for Interpretation Without and With CAD Scheme

Radiologist	A_z	
	Without CAD	With CAD
A	0.798	0.871
B	0.736	0.898
C	0.793	0.837
D	0.763	0.871
E	0.790	0.861
F	0.833	0.844
G	0.706	0.874
H	0.695	0.812
I	0.826	0.881
J	0.823	0.840
K	0.768	0.819
L	0.840	0.883
M	0.849	0.857
N	0.781	0.826
O	0.807	0.835
P	0.757	0.833
Mean	0.785	0.853

Note.—The difference for values without and with CAD scheme was statistically significant with a p value of 0.016. CAD = computer-aided diagnosis.

for solid nodules). Table 1 shows the A_z values without and with the CAD scheme for each radiologist. The average A_z value for the 16 radiologists was improved from 0.785 to 0.853 (from 0.812 to 0.892 for nodules with pure ground-glass opacity, from 0.819 to 0.863 for nodules with mixed ground-glass opacity, and from 0.784 to 0.844 for solid nodules) by a statistically significant level ($p = 0.016$) with the aid of the CAD scheme. The average ROC curves for the performance of the computer alone and the overall performance of the 16 radiologists without and with the CAD scheme for distinction between malignant and benign nodules are shown in Figure 2. The radiologists' diagnostic performance with the CAD scheme was more accurate than that of the CAD scheme alone ($p = 0.0005$). The A_z value for the CAD scheme was also greater than that of the radiologists alone ($p = 0.00006$).

Figure 3 shows the correlation between the computer outputs and the average radiologists' ratings without (Fig. 3A) and with (Fig. 3B) the CAD scheme for indicating the malignancy and benignancy of lung nodules. The radiologists' interpretations with the computer aid were, in general, more accurate than those of the radiologists alone for most of the malignant and benign nodules (Fig. 1). Note, however, that there were some cases for which the radiologists' ratings without CAD scheme were correct and the likelihood of malignancy in the computer output was incorrect. In those cases, the radiologists gave the correct ratings with the CAD scheme, as illustrated by three cancer cases (black circles) in the upper left quadrant and three benign cases (white circles) in the lower right quadrant in Figure 3B. Sample cases are shown in Figure 4.

The effect of the computer output on the average change in rating score due to the CAD is illustrated in Figure 5. The relationship between the likelihood of malignancy and the average change in confidence level (average change in ratings from without to with CAD) for each nodule by the 16 radiologists has a large correlation coefficient ($r = 0.927$). The radiologists increased their confidence level toward malignancy when the likelihood of malignancy was greater than 0.50 and decreased the confidence level toward benignancy when the likelihood measure was less than 0.50 for most of the malignant and benign nodules.

For the four clinical actions—return to annual screening, follow-up in 6 months, follow-up in 3 months, or biopsy or surgery, we attempted to quantify the changes in clinical action that were due to the CAD scheme. For malignant nodules, the average number of nodules for which clinical actions were changed by the 16 radiologists toward a beneficial effect (step up) (mean, 4.1 nodules) was greater than that toward a detrimental effect (step down) (mean, 1.2 nodules) ($p = 0.003$). For benign nodules, the number of nodules affected by the CAD scheme toward a beneficial effect (step down) and detrimental effect (step up) was 3.1 and 2.1, respectively ($p = 0.15$). Table 2 shows only the cases for which the clinical action was changed to or from the two extreme situations—that is, from biopsy or surgery to screening and from screening to biopsy or surgery. For malignant nodules, the difference was statistically significant between the change to (1.9 cases) and the change

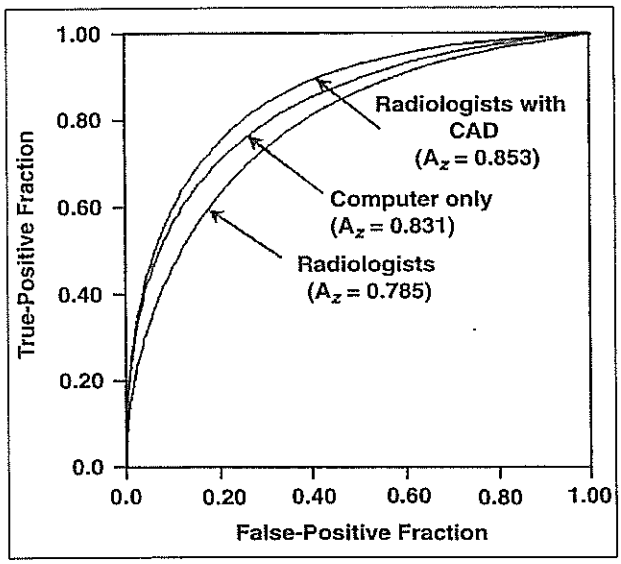


Fig. 2.—Graph shows receiver operating characteristic (ROC) curves for performance of computer alone and average performance of 16 radiologists without and with computer-aided diagnosis (CAD) scheme. Note that difference was statistically significant between radiologists without and with CAD scheme ($p = 0.016$), between computer alone and radiologists' performance without ($p = 0.00006$), and between computer alone and radiologists' performance with CAD scheme ($p = 0.0005$).

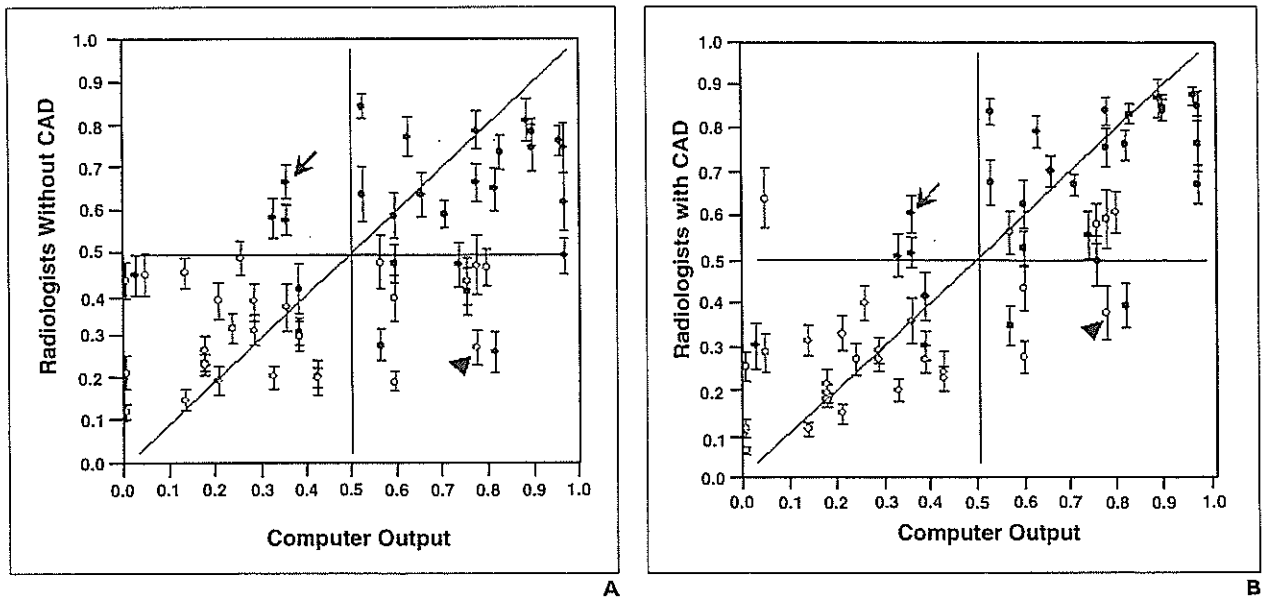


Fig. 3.—Graphs show correlation between computer output and average radiologists' ratings without and with computer-aided diagnosis (CAD) scheme for indicating likelihood of malignancy for lung nodules. ● = average radiologists' ratings for malignant nodules, ○ = average radiologists' ratings for benign nodules, horizontal lines = range of radiologists' ratings for each nodule.
A and B, Graphs show correlation between computer outputs and average radiologists' ratings without CAD (A) ($r = 0.514$) and with CAD (B) ($r = 0.784$). Note that radiologists' ratings without CAD scheme in some malignant (*upper left quadrant*) and benign (*lower right quadrant*) nodules were obviously correct, whereas likelihood of malignancy based on computer outputs alone was incorrect; even with incorrect CAD outputs, radiologists retained correct ratings. One malignant case (*arrow*) and one benign case (*arrowhead*) shown here are illustrated in Figure 4.

from (0.8 cases) biopsy or surgery ($p = 0.007$) and between the change from (0.7 cases) and the change to (0.1 cases) screening ($p = 0.02$). For benign nodules, there was no statistically significant difference between them.

Discussion

Evaluation of specific morphologic features of solitary pulmonary nodules on CT, particularly on HRCT, can help radiologists in differentiating benign from malignant lesions [11–16]. Zwirewich et al. [12] reported that increased nodule size and the presence of coarse spiculation, lobulation, and inhomogeneous central attenuation were observed with significantly greater frequency among malignant lesions, which generally appeared as solid nodules on HRCT. However, CT screening frequently detected a number of early peripheral lung adenocarcinomas, and these cancers generally appeared as nodules with pure and mixed ground-glass opacity on diagnostic HRCT [14, 15]. Some benign lesions such as nodular fibrosis also showed an HRCT pattern similar to that of adenocarcinomas and appeared as mixed ground-glass opacity nodules with a spiculated margin [16]. In this observer study, the benign lung nodules were matched in size and pattern to the malignant lung nodules, including those with pure ground-glass opacity, mixed ground-glass opacity, and solid opacity. We believe that the differential diagnosis of both benign and malignant pulmonary nodules similar in size and pattern can be difficult, and it is important to verify that a CAD scheme can assist radiolo-

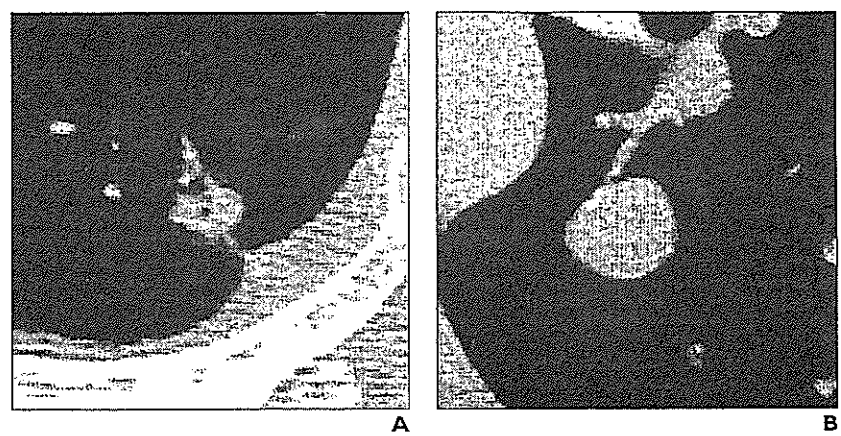


Fig. 4.—High-resolution CT (HRCT) scans show one malignant case and one benign case. Note that radiologists' interpretations without computer-aided diagnosis (CAD) scheme were correct in these cases, whereas likelihoods of malignancy based on computer outputs only were obviously incorrect; even with incorrect CAD outputs, radiologists retained correct ratings.
A, HRCT scan shows malignant lung nodule in 68-year-old man. Computer output was 0.36; radiologists' ratings without CAD, 0.67; and radiologists' ratings with CAD, 0.61.
B, HRCT scan shows benign lung nodule in 35-year-old woman. Computer output was 0.78; radiologists' ratings without CAD, 0.27; and radiologists' ratings with CAD, 0.38.

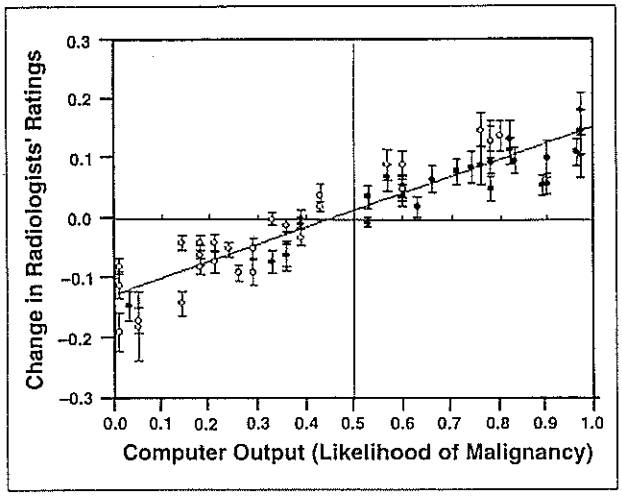


Fig. 5.—Graph shows correlation (r = 0.925) between likelihood of malignancy and average change in confidence level (rating scores) for each nodule by 16 radiologists. Malignant and benign nodules are marked by black circles and white circles, respectively. ● = average change in confidence level for malignant nodules, ○ = average change in confidence level for benign nodules, horizontal lines = range of radiologists' ratings for each nodule. Note that radiologists increased their confidence level when likelihood of malignancy was greater than 0.50 and decreased their confidence level when likelihood was less than 0.50 for most malignant and benign nodules.

TABLE 2 Cases in Which Important Clinical Actions Related to Biopsy or Screening Were Changed by 16 Radiologists as a Result of Computer-Aided Diagnosis (CAD) Scheme

Radiologist	Malignant Nodules		Benign Nodules	
	Beneficial Effect (to biopsy/from screening)	Detrimental Effect (from biopsy/to screening)	Beneficial Effect (from biopsy/to screening)	Detrimental Effect (to biopsy/from screening)
A	0/0	0/0	1/0	0/0
B	5/3	0/0	0/2	0/1
C	2/0	2/0	3/1	1/5
D	0/2	0/0	0/1	0/1
E	0/1	0/0	2/0	0/2
F	1/1	0/1	1/0	3/0
G	2/2	1/0	2/1	2/2
H	5/1	2/0	3/0	1/1
I	3/0	0/0	2/1	1/2
J	0/0	0/0	0/0	1/0
K	3/1	0/0	0/1	0/1
L	4/0	4/0	2/1	0/0
M	0/0	0/0	0/0	0/0
N	1/0	0/0	0/0	0/0
O	1/0	0/0	0/0	0/0
P	4/0	3/0	1/1	1/0
Mean (± SD)	1.9 ± 1.8[a] / 0.7 ± 0.9[b]	0.8 ± 1.3[a] / 0.1 ± 0.3[b]	1.1 ± 1.1 / 0.6 ± 0.6	0.6 ± 0.9 / 0.9 ± 1 .3

Note.—Clinical actions included return for annual screening, follow-up in 6 months, follow-up in 3 months, and biopsy or surgery. The first entry for radiologist B "5/3" means the following: five cases that had been classified as annual screening, 6-month follow-up, 3-month follow-up, or biopsy or surgery were reclassified as biopsy on the basis of the CAD output, and three cases that had been classified as screening were reclassified to 6-month follow-up, 3-month follow-up, or biopsy or surgery on the basis of CAD output.

[a]The difference for all radiologists was statistically significant with a p value of 0.007 between beneficial effect and detrimental effect among malignant lesions.

[b]The difference for all radiologists was statistically significant with a p value of 0.02 between beneficial effect and detrimental effect among malignant lesions.

gists in distinguishing these benign from malignant nodules on HRCT.

Previous studies indicated several methods for determining the probability of malignancy in masses on mammography [17, 18] and solitary pulmonary nodules on chest radiography [19–22] and chest CT [7, 23–25]. Automated feature-extraction techniques have been applied in CAD schemes for classification of malignant and benign masses on breast and lung images [7, 17, 18, 22]. Several observer studies indicated that the likelihood-of-malignancy measures can improve radiologists' diagnostic accuracy in distinguishing benign from malignant lesions on radiographs [17, 18, 23, 26] and low-dose CT scans [27]. A recent study indicated that the use of an artificial neural network (ANN) as a computer aid based on attending radiologists' subjective rating scores improved radiologists' performance in terms of A_z value from 0.831 to 0.959 in differentiating benign from malignant pulmonary nodules on HRCT [25]. The performance of our automated feature-extraction scheme for all nodules in our database (A_z = 0.937) was comparable to that of the ANN by use of subjective ratings (A_z = 0.951) [25]. Our observer study indicates the usefulness of our automated computerized scheme in the classification of pulmonary nodules on HRCT images. In the future, therefore, an automated computerized scheme as second opinion may be acceptable to radiologists in clinical situations.

Our automated computerized scheme is based on various objective features (size, contrast, shape, margin, internal opacity, and internal features) of the nodules. The performance of the CAD scheme was evaluated on the basis of a leave-one-out testing method using 61 malignant and 183 benign nodules. In the computer output, a misclassification by the CAD system was observed to occur in large benign solid nodules (Fig. 4B) and in nodules with mixed ground-glass opacity, including benign (Fig. 1D) and malignant lesions (Fig. 4A). These misclassifications probably occurred because our database was obtained from a CT screening program in which all (15 lesions) solid malignant lesions were more than 10 mm, 94% (133/141) of solid benign nodules were 10 mm or less, and in a nodule with mixed ground-glass opacity, it was more difficult to differentiate benign from malignant by the CAD scheme. Also, there was a limitation in this observer study because the 56 nodules were included for developing the CAD scheme. The number of nodules, especially malignant nodules in our database, was not enough to divide training and test groups in this study,

and we plan to use an independent database from other CT screening programs to test the usefulness of our CAD scheme in the future.

Our results in this study showed that the radiologists' performance with CAD scheme (0.853) was greater than that of either radiologists alone (0.785) or computer output alone (0.831), with statistically significant differences in A_z values. The radiologists generally increased or decreased their confidence level when the likelihood of malignancy was above or below 0.50, respectively, and the changes based on CAD output for most nodules were toward a beneficial effect. Important findings are that the radiologists' initial ratings without CAD were clearly correct for some nodules and that even when the computer output indicated incorrect results, no serious detrimental effect to the radiologists' ratings as a result of the CAD output occurred. Thus, radiologists were able to maintain their correct judgments when nodules appeared obviously benign or malignant despite an incorrect CAD output. In addition, the correct computer output was able to assist radiologists in improving their decisions on many subtle cases. Therefore, this study indicated that a synergistic improvement in observers' interpretation by use of a CAD scheme as a second opinion was possible, because the radiologists were able to maintain their own correct opinions on some obvious cases, whereas the computer output assisted in improving their decisions on the majority of subtle cases.

In this study, we quantified the changes due to the CAD scheme in two extreme situations—that is, changes to or from biopsy or screening, which are important decisions in cancer screening. The results indicate the benefit of the computer aid to radiologists in making correct recommendations for malignant lesions. However, no significant benefit of the computer aid to radiologists was observed for benign nodules. Possible reasons might be that because this study was based on lung cancer CT screening, radiologists were highly alerted to avoid making underinterpretations for subtle pulmonary nodules regardless of the result of the CAD scheme.

Acknowledgments

We thank Takeshi Kobayashi, Kazuto Ashizawa, Naohiro Matsuyama, Hajime Abiru, Tetsuji Yamaguchi, Chaotong Zhang, Peter MacEneaney, Ulrich Bick, Christopher Straus, Edward Michale, Gregory Scott Stacy, Akiko Egawa, Tomoaki Okimoto, Kazunori Minami, and Shuji Sakai, for participating as observers; Shigehiko Katsuragawa, for helpful suggestions; and Elisabeth Lanzl for improving the manuscript.

References

1. 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
2. Sone S, Takashima S, Li F, et al. Mass screening for lung cancer with mobile spiral computed tomography scanner. *Lancet* 1998;351:1242–1245
3. Henschke CI, MacCauley DI, Yankelevitz DF, et al. Early lung cancer action project: overall design and findings from baseline screening. *Lancet* 1999;354:99–105
4. Diederich S, Wormanns D, Semik M, et al. Screening for early lung cancer with low-dose spiral CT: prevalence in 817 asymptomatic smokers. *Radiology* 2002;222:773–781
5. Swensen SJ, Jett JR, Sloan JA, et al. Screening for lung cancer with low-dose spiral computed tomography. *Am J Respir Crit Care Med* 2002;165:508–513
6. Mahadevia PJ, Fleisher LA, Frick KD, Eng J, Goodman SN, Powe NR. Lung cancer screening with helical computed tomography in older adult smokers. *JAMA* 2003;289:313–322
7. Aoyama M, Li Q, Katsuragawa S, Li F, Sone S, Doi K. Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images. *Med Phys* 2003;30:387–394
8. Metz CE, Herman BA, Shen JH. Maximum-likelihood estimation of receiver operating (ROC) curves from continuously distributed data. *Stat Med* 1998;17:1033–1053
9. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jackknife method. *Invest Radiol* 1992;27:723–731
10. Metz CE. Quantification of failure to demonstrate statistical significance: the usefulness of confidence intervals. *Invest Radiol* 1993;28:59–63
11. Erasmus JJ, Connolly JE, McAdams HP, Roggli VL. Solitary pulmonary nodules. I. Morphologic evaluation for differentiation of benign and malignant lesions. *RadioGraphics* 2000;20:43–58
12. Zwirewich CV, Vedal S, Miller RR, Müller NL. Solitary pulmonary nodule: high-resolution CT and radiologic-pathologic correlation. *Radiology* 1991;179:469–476
13. Takashima S, Sone S, Li F, et al. Small solitary pulmonary nodules (≤ 1 cm) detected at population-based CT screening for lung cancer: reliable high-resolution CT features of benign lesions. *AJR* 2003;180:955–964
14. Yang Z-G, Sone S, Takashima S, et al. High-resolution CT analysis of small peripheral lung adenocarcinomas revealed on screening helical CT. *AJR* 2001;176:1399–1407
15. Henschke CI, Yankelevitz DF, Mirtcheva R, et al. CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules. *AJR* 2002;178:1053–1057
16. Li F, Sone S, Maruyama Y, et al. Correlation between high-resolution computed tomographic, magnetic resonance and pathological findings in cases with non-cancerous but suspicious lung nodules. *Eur Radiol* 2000;10:1782–1791
17. Chan HP, Sahiner B, Helvie MA, et al. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. *Radiology* 1999;212:817–827
18. Huo Z, Giger ML, Vybomy CJ, Metz CE. Breast cancer: effectiveness of computer-aided diagnosis observer study with independent database of mammograms. *Radiology* 2002;224:560–568
19. Gurney JW. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis. I. Theory. *Radiology* 1993;186:405–413
20. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules: application to small radiologically indeterminate nodules. *Arch Inter Med* 1997;157:849–855
21. Nakamura K, Yoshida H, Engelmann R, et al. Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks. *Radiology* 2000;214:823–830
22. Aoyama M, Li Q, Katsuragawa S, MacMahon H, Doi K. Automated computerized scheme for distinction between benign and malignant solitary pulmonary nodules on chest images. *Med Phys* 2002;29:701–708
23. Henschke CI, Yankelevitz DF, Mateescu I, Brettle DW, Rainey TG, Weingard FS. Neural networks for the analysis of small pulmonary nodules. *Clin Imaging* 1997;21:390–399
24. McNitt-Gray MF, Hart EM, Wyckoff N, Sayre JW, Goldin JG, Aberle D. A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results. *Med Phys* 1999;26:880–888
25. Matsuki Y, Nakamura K, Watanabe H, et al. Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: evaluation with receiver operating characteristic analysis. *AJR* 2002;178:657–663
26. Shiraishi J, Abe H, Engelmann R, Aoyama M, MacMahon H, Doi K. Computer-aided diagnosis to distinguish benign from malignant solitary pulmonary nodules on radiographs: ROC analysis of radiologists' performance—initial experience. *Radiology* 2003;227:469–474
27. Li Q, Li F, Shiraishi J, Katsuragawa S, Sone S, Doi K. Investigation of new psychophysical measures for evaluation of similar images on thoracic CT for distinction between benign and malignant nodules. *Med Phys* 2003;30:2584–2593

International Congress Series 1268 (2004) 946–951

ELSEVIER

www.ics-elsevier.com

Usefulness of computerized scheme for differentiating benign from malignant lung nodules on high-resolution CT

Feng Li[a,*], Qiang Li[a], Masahito Aoyama[b], Junji Shiraishi[a], Hiroyuki Abe[a], Kenji Suzuki[a], Roger Engelmann[a], Shusuke Sone[c], Heber MacMahon[a], Kunio Doi[a]

[a] *Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, The University of Chicago, MC-2026, 5841 South Maryland Avenue, Chicago, IL 60637, USA*
[b] *Department of Intelligent Systems, Faculty of Information Sciences, Hiroshima City University, Hiroshima 731-3194, Japan*
[c] *Azumi General Hospital, Ikeda, Nagano 399-8695, Japan*

Abstract. A computer-aided diagnosis (CAD) scheme for determination of the likelihood of malignancy of 244 nodules on high-resolution CT (HRCT) was developed. The performance (A_z) for 16 radiologists was improved from 0.785 to 0.853 ($P=0.02$) with the aid of the CAD scheme by use of 56 nodules, including 28 cancerous and 28 benign nodules which were matched in size and pattern to the cancers. Our purpose in this study was to investigate further whether a CAD scheme can assist radiologists in distinguishing benign from malignant nodules in different groups. The results indicated that A_z values for radiologists without and with the CAD scheme were improved from 0.770 to 0.855 for general radiologists ($P=0.01$) and from 0.805 to 0.850 for chest radiologists ($P=0.12$); from 0.717 to 0.821 for nodules at 6–10 mm ($P=0.04$) and from 0.837 to 0.901 for nodules at 11–20 mm ($P=0.04$); and from 0.812 to 0.892 for nodules with pure ground-glass opacity (GGO) ($P=0.149$), from 0.819 to 0.863 for nodules with mixed GGO ($P=0.196$), and from 0.784 to 0.844 for solid nodules ($P=0.334$). CAD has the potential to improve the diagnostic accuracy in distinguishing benign nodules from malignant ones in different groups on HRCT.

Keywords: Lung nodule; Computer-aided diagnosis (CAD); Receiver operating characteristic (ROC) curve

1. Introduction

Low-dose spiral CT has been applied for cancer screening and led to early detection of lung cancer in some countries [1–3]. However, simultaneous or additional diagnostic high-resolution CT (HRCT) was needed for distinction between early cancers, most of

* Corresponding author. Tel.: +1-773-834-5093; fax: +1-773-702-0371.
E-mail address: fli@kurt.bsd.uchicago.edu (F. Li).

0531-5131/
doi:10.1016/j.ics.2004.03.083

which were nodules with ground-glass opacity (GGO), and a large number of benign nodules, which were detected as suspicious lesions by screening with low-dose CT [4,5]. It is important to differentiate benign from malignant nodules on the first diagnostic HRCT in order to reduce the number of follow-up HRCT examinations. The database used in this study was obtained as part of follow-up diagnostic work in a lung cancer screening program in Japan [1]. The database consisted of 244 small, non-calcified (3–20 mm) nodules, including nodules with pure GGO, mixed GGO, and solid opacity on the HRCT (1 mm collimation). We found that certain HRCT characteristics, which were reviewed subjectively by radiologists, could be helpful in differentiating small malignant from benign nodules [5]. Recently, we developed an automated computerized scheme for distinction between malignant and benign lesions. Our observer study indicated that the average A_z value for 16 radiologists was improved from 0.785 to 0.853 by a statistically significant amount ($P=0.02$) with the aid of a computer-aided diagnosis (CAD) scheme for 28 malignant nodules and 28 benign nodules, which were matched in size and pattern to the cancers [6]. In this study, we further investigated whether a CAD scheme could assist radiologists in distinguishing benign from malignant nodules in different groups on HRCT.

2. Methods

2.1. Database

The HRCT database used in this study consisted of 61 malignant (mean 12 mm; range, 6–19 mm) and 183 benign nodules (mean 7 mm; range, 3–20 mm). The patterns of malignant and benign nodules included pure GGO (malignant vs. benign: 18 vs. 12), mixed GGO (28 vs. 30) and solid opacity (15 vs. 141), which were subjectively reviewed by three radiologists. The mean size of all solid malignant lesions was more than 10 mm, and the mean size of 133 (94%) solid benign nodules was 10 mm or less.

2.2. Computer-aided diagnosis

We developed an automated computerized scheme for determination of the likelihood of malignancy of lung nodules, which was based on various objective features (size, contrast, shape, margin, internal opacity, and internal features) of the nodules. The CAD scheme was based on multiple slices of HRCT derived from 2D and 3D volume data by use of linear discriminant analysis. The performance of our CAD scheme yielded an A_z value of 0.937 in the distinction between malignant and benign nodules for all three patterns. The A_z values of the classification scheme were 0.919 for nodules with pure GGO, 0.852 for nodules with mixed GGO, and 0.957 for solid nodules.

2.3. Observer study

The cases used in this observer study consisted of 28 malignant nodules, which were randomly selected from 61 primary lung cancers, and 28 benign nodules, which were

Table 1
A_z values for 16 radiologists for interpretation without and with CAD scheme

Radiologist	Without CAD scheme	With CAD scheme
Chest		
A	0.763	0.871
B	0.833	0.844
C	0.823	0.840
D	0.768	0.819
E	0.840	0.883
F	0.849	0.857
G	0.757	0.833
Mean	0.805	0.850
General		
H	0.798	0.871
I	0.736	0.898
J	0.793	0.837
K	0.790	0.861
L	0.706	0.874
M	0.695	0.812
N	0.826	0.881
O	0.781	0.826
P	0.807	0.835
Mean	0.770	0.855
Mean for all	0.785	0.853

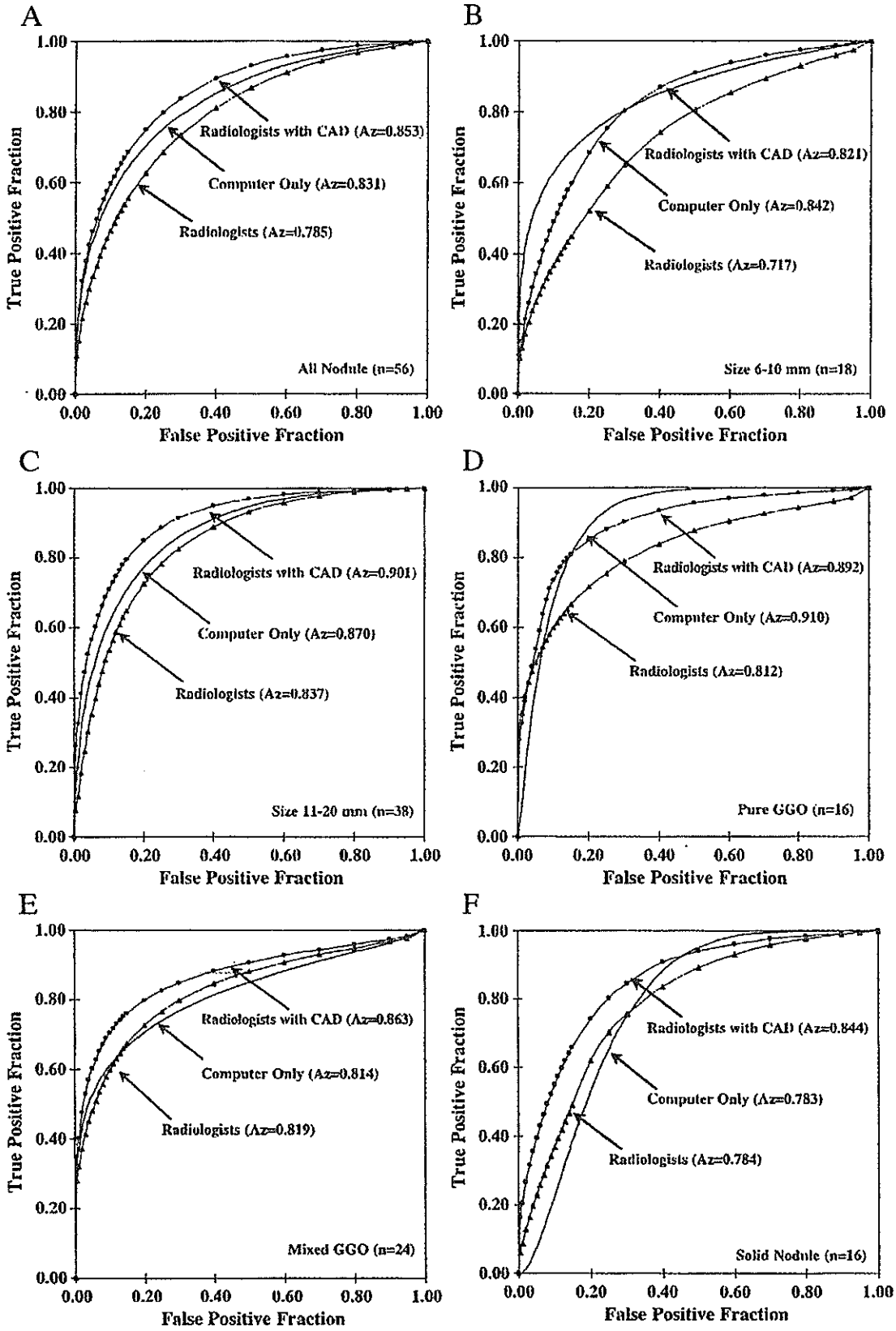
The difference was statistically significant with a p value 0.01 for general radiologists (0.12 for chest radiologists) between radiologist without and with CAD scheme. No statistically significant difference was found between chest and general radiologists for each condition without and with CAD scheme.

selected from 183 benign nodules matched in size and pattern to the cancers. Sixteen radiologists participated in this study.

A number of axial images for each nodule on HRCT were displayed for interpretation in cine mode on a CRT monitor. For a training session before the test, we provided five different cases so that the observers could learn how to operate the cine mode interface and how to take into account the computer output in their decision. We informed observers that the sensitivity and specificity of our CAD scheme, using a threshold of 0.50 (50%) for the likelihood of malignancy, were 80% and 75%, respectively. The images were presented to radiologists, first without and then with the computer output, who were asked to indicate their confidence level regarding the malignancy of a nodule.

The performance (A_z) of the CAD scheme was 0.831 in the distinction between the 28 malignant and 28 benign nodules used in this observer study. The average A_z value for the 16 radiologists was improved from 0.785 to 0.853 by a statistically significant level ($P=0.02$) with the aid of the CAD scheme for the 56 nodules.

Fig. 1. ROC curves show the performance of the computer alone and the average performance of 16 radiologists without and with CAD scheme for two size and three pattern groups. (A) All nodules, (B) small nodules, (C) large nodules, (D) pure GGO, (E) mixed GGO, (F) solid nodules.


A
True Positive Fraction
False Positive Fraction
Radiologists with CAD (Az=0.853)
Computer Only (Az=0.831)
Radiologists (Az=0.785)
All Nodule (n=56)
B
Radiologists with CAD (Az=0.821)
Computer Only (Az=0.842)
Radiologists (Az=0.717)
Size 6-10 mm (n=18)
C
Radiologists with CAD (Az=0.901)
Computer Only (Az=0.870)
Radiologists (Az=0.837)
Size 11-20 mm (n=38)
D
Radiologists with CAD (Az=0.892)
Computer Only (Az=0.910)
Radiologists (Az=0.812)
Pure GGO (n=16)
E
Radiologists with CAD (Az=0.863)
Computer Only (Az=0.814)
Radiologists (Az=0.819)
Mixed GGO (n=24)
F
Radiologists with CAD (Az=0.844)
Computer Only (Az=0.783)
Radiologists (Az=0.784)
Solid Nodule (n=16)

2.4. Different groups

The 16 radiologists were divided into two groups, which included seven chest radiologists with the mean experience of 15 years and nine general radiologists with the mean experience of 13 years. The 56 nodules included two groups of nodules of different size (9 small nodules of 6–10 mm and 19 large nodules of 11–20 mm), and also three groups with different patterns (8 nodules with pure GGO, 12 nodules with mixed GGO, and 8 solid nodules) for both malignant and benign lesions.

2.5. Data analysis

The confidence level ratings by each observer were analyzed by use of receiver operating characteristic (ROC) methodology, and a quasi-maximum likelihood estimation of the binormal distribution was fitted to the radiologists' confidence ratings [7]. The statistical significance of the difference in A_z values between observer readings without and with the CAD scheme was tested by use of the Dorfman–Berbaum–Metz method [8], which included both reader variation, and case sample variation by means of an analysis-of-variance approach.

3. Results

The A_z values for radiologists without and with the CAD scheme were improved from 0.770 to 0.855 for the nine general radiologists ($P=0.01$), and from 0.805 to 0.850 for the seven chest radiologists ($P=0.12$). A_z values without and with the CAD scheme for each radiologist are listed in Table 1.

The A_z values of the classification CAD scheme were 0.842 for small nodules, 0.870 for large nodules, 0.910 for nodules with pure GGO, 0.814 for nodules with mixed GGO, and 0.783 for solid nodules. The A_z values for radiologists without and with the CAD scheme were improved from 0.717 to 0.821 for small nodules ($P=0.04$) and from 0.837 to 0.901 for large nodules ($P=0.04$); and from 0.812 to 0.892 for nodules with pure GGO ($P=0.149$), from 0.819 to 0.863 for nodules with mixed GGO ($P=0.196$), and from 0.784 to 0.844 for solid nodules ($P=0.334$). ROC curves for use of the computer alone and the average performance of 16 radiologists without and with the CAD scheme for distinction between malignant and benign nodules are shown in Fig. 1 (two size and three pattern groups).

4. Conclusion

CAD has the potential to improve the diagnostic accuracy in distinguishing benign nodules from malignant ones in different groups on HRCT.

Acknowledgements

The authors are grateful to Takeshi Kobayashi, MD, Kazuto Ashizawa, MD, Naohiro Matsuyama, MD, Hajime Abiru, MD, Tetsuji Yamaguchi MD, Chaotong Zhang, MD, Peter MacEneaney, MD, Ulrich Bick, MD, Christopher Straus, MD, Edward Michals, MD, Gregory Scott Stacy, MD, Akiko Egawa, MD, Tomoaki Okimoto, MD, Kazunori Minami,

MD, and Shuji Sakai, MD, for participating as observers; and to Shigehiko Katsuragawa, PhD, for helpful suggestions. This work was partly supported by USPHS grants CA62625 and CA 98119. H. MacMahon and K. Doi are shareholders of R2 Technology, Los Altos, CA. K. Doi is a shareholder of Deus Technology, Rockville, MD.

References

[1] S. Sone, et al., Mass screening for lung cancer with mobile spiral computed tomography scanner, Lancet 351 (1998) 1242–1245.
[2] C.I. Henschke, et al., Early lung cancer action project: overall design and findings from baseline screening, Lancet 354 (1999) 99–105.
[3] S. Diederich, et al., Screening for early lung cancer with low-dose spiral CT: prevalence in 817 asymptomatic smokers, Radiology 222 (2002) 773–781.
[4] C.I. Henschke, et al., CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules, AJR 178 (2002) 1053–1057.
[5] F. Li, Comparison of thin section CT findings in malignant and benign nodules in CT screening for lung cancer, Radiology (in press).
[6] F. Li, et al., Improvement in radiologists' performance for differentiating small benign from malignant lung nodules on high-resolution CT by using computer-estimated likelihood of malignancy, AJR (in press).
[7] C.E. Metz, B.A. Herman, J.H. Shen, Maximum-likelihood estimation of receiver operating (ROC) curves from continuously distributed data, Stat. Med. 17 (1998) 1033–1053.
[8] D.D. Dorfman, K.S. Berbaum, C.E. Metz, ROC rating analysis: generalization to the population of readers and cases with the jackknife method, Invest. Radiol. 27 (1992) 723–731.

Effect of Temporal Subtraction Images on Radiologists' Detection of Lung Cancer on CT: Results of the Observer Performance Study with Use of Film Computed Tomography Images[1]

Hiroyuki Abe, MD, PhD, Takayuki Ishida, PhD, Junji Shiraishi, PhD, Feng Li, MD, PhD, Shigehiko Katsuragawa, PhD, Shusuke Sone, MD, Heber MacMahon, MD, Kunio Doi, PhD

Rationale and Objectives. To evaluate the effect of temporal subtraction images on the radiologists' detection of early primary lung cancer in computed tomography (CT) scans.

Materials and Methods. Fourteen cases with primary lung cancer and 16 normal cases were used for this study from a database of low-dose CT images, which were obtained from a lung cancer screening program in Nagano, Japan. Images were obtained with a single-detector helical CT scanner using 10 mm collimation and 2:1 pitch. Each case had both previous and current CT scans. Temporal subtraction images were obtained by subtracting the warped previous images from the current images. Seven radiologists, including four attendings and three residents, provided their confidence levels for the presence or absence of lung cancers with use of film CT images without and with temporal subtraction images. Receiver operating characteristic analysis was used to compare their performance without and with temporal subtraction images.

Results. The mean Az values (area under the receiver operating characteristic curve) of seven observers without and with temporal subtraction images were 0.868 and 0.930, respectively. Diagnostic accuracy was significantly improved by using temporal subtraction images ($P = .007$). Temporal subtraction images were especially useful when a nodule was present near the pulmonary hilum, where radiologists tended to overlook it.

Conclusion. The temporal subtraction technique can significantly improve the sensitivity and specificity for detection of lung cancer on CT scans.

Key Words. Lung cancer; temporal subtraction; computed tomography (CT); computer-aided diagnosis (CAD); observer study.

Acad Radiol 2004; 11:1337–1343

[1] From Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, The University of Chicago, MC 2026, 5841 S Maryland Ave, Chicago, IL 60637 (H.A., J.S., F.L., H.M., K.D.); Hiroshima International University, Hiroshima, Japan (T.I.); Kumamoto University, Kumamoto, Japan (S.K.); and Azumi General Hospital, Nagano, Japan (S.S.). Received May 16, 2004; revision requested June 24; received July 6; revision requested August 10; received August 27; accepted August 29. Supported by US Public Health Service grant no. CA62625. Address correspondence to H.A. e-mail: habe@uchicago.edu

doi:10.1016/j.acra.2004.08.010

Screening for lung cancer with low-dose computed tomography (CT) has been become popular in the United States and Japan, partly because earlier-stage cancer can be detected with CT than with chest radiographs (1,2). However, radiologists sometimes overlook lung cancer even with CT, despite the fact that CT has high spatial and contrast resolution (3–5).

A temporal subtraction technique for chest radiographs, in which areas of interval change were enhanced by using the previous radiograph as a subtraction mask, has been developed previously (6–8). The observer study confirmed that the temporal subtraction technique was helpful

to radiologists in the detection of subtle nodules which might have been missed without it (9–11). To our knowledge, there has been no report published regarding the use of temporal subtraction for CT. The purpose of this study is to report the results of an observer study conducted to determine if a temporal subtraction technique could be useful to improve the accuracy of radiologists in detecting subtle nodules caused by lung cancer on CT.

MATERIALS AND METHODS

Cases

We selected 14 cases with primary lung cancer and 16 normal cases for this study from a database of low-dose CT images, which were obtained from a CT screening program for lung cancer conducted in Nagano, Japan, from 1996 to 1999. The database consists of 7,847 screenees including 87 cases with primary lung cancer. In the screening program, all subjects gave informed consent. All cases used in this study had at least two sequential CT images with from 1- to 2-year interval between them. All of the CT images were obtained with a single detector helical CT scanner (W950SR; Hitachi Medical, Tokyo, Japan) with a 10 mm collimation and one rotation of the x-ray tube per 2-seconds with table speed of 10 mm/sec (pitch; 2:1). All of the scans were performed in a 32-second breath hold with scanning from below the level of diaphragm to above the level of apex. All of the images were reconstructed as 10 mm slice thickness. Although the number of slices obtained in these cases was 32, all of cases used in this study covered the entire lungs within 30 slices.

The selection criteria for positive cases included solitary, primary lung cancer and cases having at least two sequential CT scans with growth of a nodule observed in the latter CT. The size of the nodules ranged from 5–24 mm (mean, 12 mm). The normal cases were selected randomly without the knowledge of the quality of subtraction images produced. All of the normal cases were verified by three experienced radiologists (H.A, F.L, H.M), who did not participate in the observer study, for the absence of nodules or nodule-like lesions.

The location of the tumor was categorized into three areas; peripheral lung field, central lung field, and perihilum. When the tumor was located within 2 cm from the pulmonary hilum, the location of the tumor was determined as perihilum. In the same manner, when the tumor was located in the lung field within 2 cm from the pleura, the location of the tumor was determined as peripheral lung field. (If the tumor was located within 2 cm both from the pulmonary hilum and from the pleura, the location of the tumor was determined as perihilum.) The location of the remaining tumors was determined as in central lung field.

Primary cancers consisted of adenocarcinoma (n = 9), small cell carcinoma (n = 3), and squamous cell carcinoma (n = 2). Of these cancers, eight represented ground glass opacity and six showed solid opacity. The average age of the patients was 69.6 for cases with cancer (range, 56–89), and 51.4 for normal cases (range, 44–74).

Table 1
A_z Value of Each Observer in the Detection of Lung Cancer Without and With Temporal Subtraction Image

		Without	With
Attendings	A	0.810	0.949
	B	0.861	0.932
	C	0.905	0.928
	D	0.858	0.943
	mean	0.858	0.938
Residents	E	0.933	0.965
	F	0.757	0.812
	G	0.955	0.983
	mean	0.882	0.920
All	mean	0.868	0.930

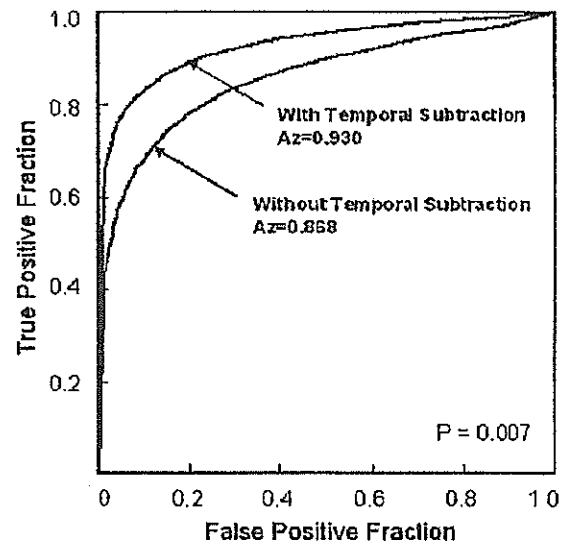


Figure 1. ROC curves for the average performance of seven observers in the detection of lung cancer on CT without and with temporal subtraction. The performance with temporal subtraction is significantly higher than that without it ($P = .007$).

Table 2
The Number of Beneficial and Detrimental Effects in Both Normal and Cancer Cases

		Beneficial effect			Detrimental effect		
		Normal	Cancer	Total	Normal	Cancer	Total
Attendings	A	11	2	13	1	1	2
	B	1	2	3	0	0	0
	C	6	1	7	1	2	3
	D	13	2	15	1	2	3
Residents	E	7	1	8	0	1	1
	F	6	2	8	0	2	2
	G	4	0	4	0	0	0
	mean	6.9	1.4	8.3	0.4	1.2	1.6

Temporal Subtraction Technique

The temporal subtraction technique on chest radiographs has been described in detail previously (6–8). In this study, we used an initial version of the temporal subtraction scheme to create the subtraction images. In the scheme, the selection of the corresponding section in the two sets of CT images and image shift correction between the current and the previous image were performed manually.

For making a subtraction of CT image in each slice, we used an iterative image warping technique. First, a number of template regions of interests (ROIs) with a 32 × 32 matrix size and the corresponding search area ROIs with a 64 × 64 matrix size were selected automatically on the previous and current images, respectively. Then, the shift values for template ROIs, which would match to the corresponding areas in search area ROIs were determined for all pairs of selected ROIs by means of a cross-correlation technique. The previous image was nonlinearly warped according to local shift vectors for "best-matching," which was determined by highest cross-correlation value (6,12–14). The warped previous image was then used for the second warping for further reduction of misregistration artifacts. Finally, the temporal subtraction image was obtained by subtracting the second warped previous image from the current image.

Observer Test and Data Analysis

Four attending radiologists (years of experience, 6–16; mean, 12.0 years) and three radiology residents (one third-year, and two fourth-year residents) participated in the observer study. Before the test, the observers were told that about half of the cases had solitary lung cancer and the rest were normal and that the benign lesions such as scar, atelectasis, and interstitial opacities should be ignored. For a training session, four cases (two cancer cases and two normal cases) were presented for each observer to become familiar with the test. In the test, observers were required to indicate their diagnostic decisions without temporal subtraction images initially, and then with temporal subtraction images, sequentially. All the slice images for previous CT, current CT, and the subtraction technique were printed separately on 14 × 17 inch films (capable of having 30 slice images maximum) with a window width of 1,500 Hounsfield unit (HU) and a window level of −700 HU.

Observers indicated their confidence levels regarding the presence or absence of cancers using an analog continuous-rating scale with a line-checking method. There were two identical 7-cm-long lines arranged one above the other, with about 5 mm distance in between. For the initial ratings, the observers marked their confidence levels along the upper line. Points toward the right or left end of the bar indicated the observer's greater confidence in a positive or negative result, respectively. Only when the second ratings were different from the initial ratings, observers indicated their confidence levels in the lower line. Test sequence is as follows; previous and current CT images were presented to observers, who were asked to indicate their confidence rating without temporal subtraction first. Immediately after the completion of initial rating, temporal subtraction images were presented together with those previous and current CT images, and observers were given the opportunities to change their confidence ratings. There was no time limit for the test. In the test, no observer requested a break, and the average time for each reading session was approximately 1 hour.

Table 3
The Number of Observers Who Detected the Cancer Without and With Temporal Subtraction, and the Size and Location of the Cancer

Case No.	Size (mm)	Location*	Detected without temporal subtraction	Missed initially, but detected with the aid of temporal subtraction
1	21	perihilum, rt	1	6
2	6	peripheral, lt	1	0
3	10	peripheral, lt	2	0
4	5	peripheral, rt.	2	0
5	8	central, lt	3	1
6	10	perihilum, rt	4	0
7	11	perihilum, lt.	5	2
8	21	central, rt	6	0
9	15	peripheral, lt	6	0

*peripheral: peripheral lung field; central: central lung field

The observer performance in the detection of lung cancer, without and with temporal subtraction, was evaluated by means of receiver operating characteristic (ROC) analysis. A binormal ROC curve was fitted to each observer's confidence rating data by using the maximum likelihood estimation (15). A computer program, LABROC5, was used for obtaining binormal ROC curves by use of the confidence level scored by measuring the distance from the left end of the line to the marked point on the continuous-rating scale and converting the measurement to a scale from 0.00 to 1.00. The A_Z value was calculated for each fitted curve. The statistical significance of the difference between the A_Z values obtained without and with the aid of temporal subtraction images was tested by use of the paired *t*-test.

We assumed that the temporal subtraction technique had a clinically relevant effect on an observer's diagnosis when there was a difference in the rating scores of 30% or more between the first and second ratings. We also assumed that the temporal subtraction technique was beneficial only when the second rating was on the point more than half of the continuous-rating scale toward the correct end, in addition to the 30% change in the confidence level. The temporal subtraction technique was detrimental when there was a 30% change toward the incorrect end of the continuous-rating scale. We assumed that the cancer was overlooked by the observer when the rating of confidence level was less than 0.5 in positive cases. The statistical significance of the difference in these results was tested by use of the paired *t*-test.

RESULTS

The results of the observer's performance without and with temporal subtraction images are shown in Table 1. All observers improved their performance in the detection of lung cancer by using temporal subtraction images. The average area under the ROC curve (A_Z value) for seven observers was significantly improved by using the temporal subtraction technique, from 0.868 to 0.930 ($P = .007$) (Fig 1). Table 2 shows the beneficial and detrimental effect of the temporal subtraction on each observer. The average number of cases for beneficial effect was significantly ($P = .006$) greater than that for detrimental effect. The beneficial effect was appreciated significantly ($P = .009$) more in normal cases than in cancer cases.

Five cancers (case No. 2–6) were not detected by at least three observers both without and with temporal subtraction (Table 3). Of these five cancers, all cancers measured within 10 mm and three cancers located in peripheral lung field. Two cancers (case No. 1, 7) were overlooked by multiple observers initially, but were detected with the aid of temporal subtraction (Table 3). Those two cancers were located at perihilum (Fig 2).

DISCUSSION

Subtle nodules and nodules overlapped with normal structures or located near the large vessels are likely to be "missed nodules" on CT (16). Because the temporal subtraction technique can remove most of the normal structures from the images, it appears appropriate to use this technique for detection of nodules which are obscured by the normal structures (10). Observers tended to overlook nodules located at perihilum without temporal subtraction in this study, in spite of the fact that radiologists might be expected to detect lesions more sensitively in an observer test than in daily clinical work, because the task in an

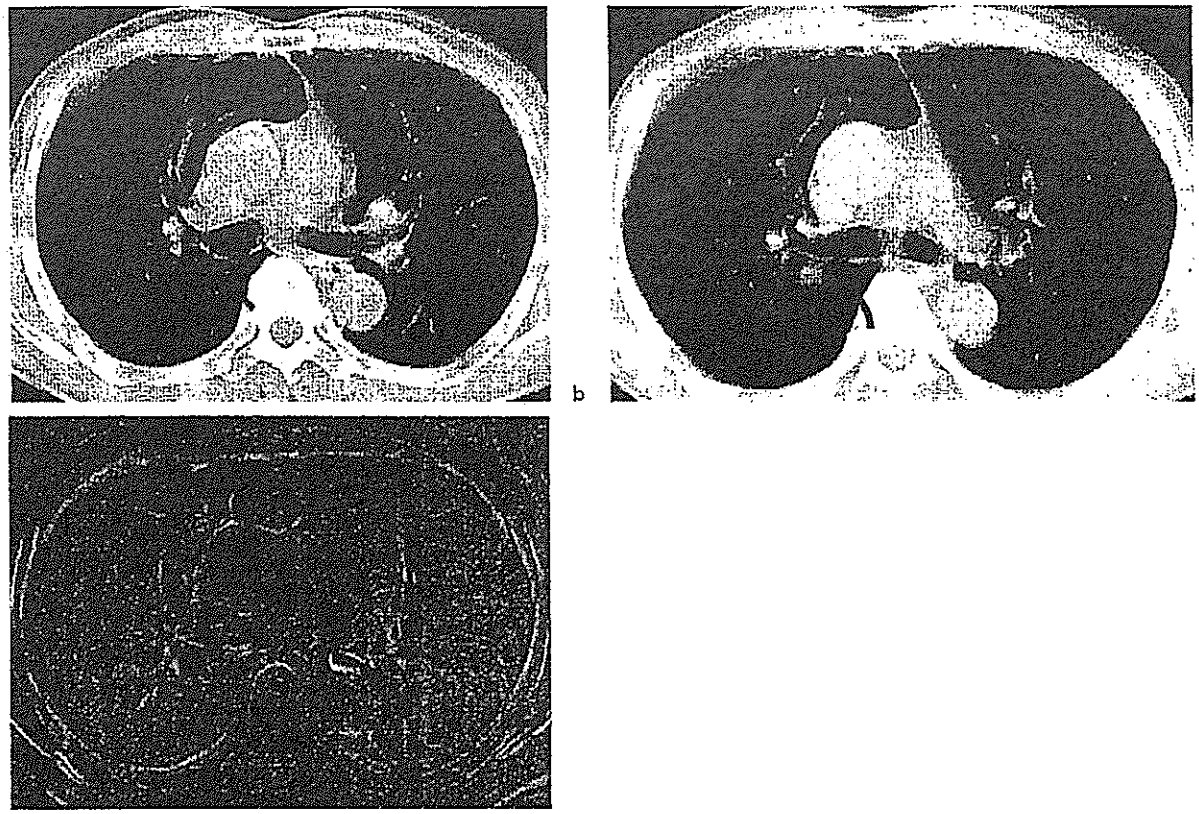


Figure 2. Case 1 (**a, previous CT; b, current CT; c, temporal subtraction image**): A lung cancer located behind the right pulmonary hilum on the current CT image (**b, curved arrow**). At the same section on the previous CT, there is only a small, faint opacity at the same location (**a, arrow**). Temporal subtraction clearly demonstrates the new nodule as a dark area that stands out from the gray background (**c**). Six of seven observers missed this lesion initially, but detected it after viewing the temporal subtraction image.

observer test is simplified and the observer is not subject to interruptions. In this study, temporal subtraction images clearly helped the radiologists not to miss cancers at least in two cases (Fig 2). On the other hand, small cancers located in the periphery tended to be overlooked even with temporal subtraction, partly because the lesions in the subtraction images were sometimes obscured by the artifacts caused by misregistration the normal structures (Fig 3). The detectability of these small cancers may be improved, if misregistration artifacts in the subtraction images can be reduced in the future. Despite these artifacts, however, only a small number of detrimental effects were found with use of the temporal subtraction. This is likely because the temporal subtraction images are only used as a guide to direct attention to suspicious area in the original images. If the opacity on the subtraction image is not confirmed on the original image, it is assumed to be an artifact. As to the beneficial effect, it tended to be greater in normal cases, to confirm the absence of new nodule, than in cancer cases. This could bring great benefit if temporal subtraction were used in a screening program where the majority of examinations are normal.

In this study, we used films, not a soft copy display, to show CT images for the observers. Although cine mode soft copy viewing of spiral CT images of the chest may improve the radiologists' ability to detect nodules compared with hard copy reading (17), we believe that the radiologists' performance would be improved with temporal subtraction even if a cine viewing were used because lung cancers was commonly overlooked even with soft copy displays (3).

The observer performance test is usually conducted in two ways (ie, an independent test and a sequential test). The independent test consists of two separate sessions to obtain results without and with aid independently. The sequential test is conducted in one session in which

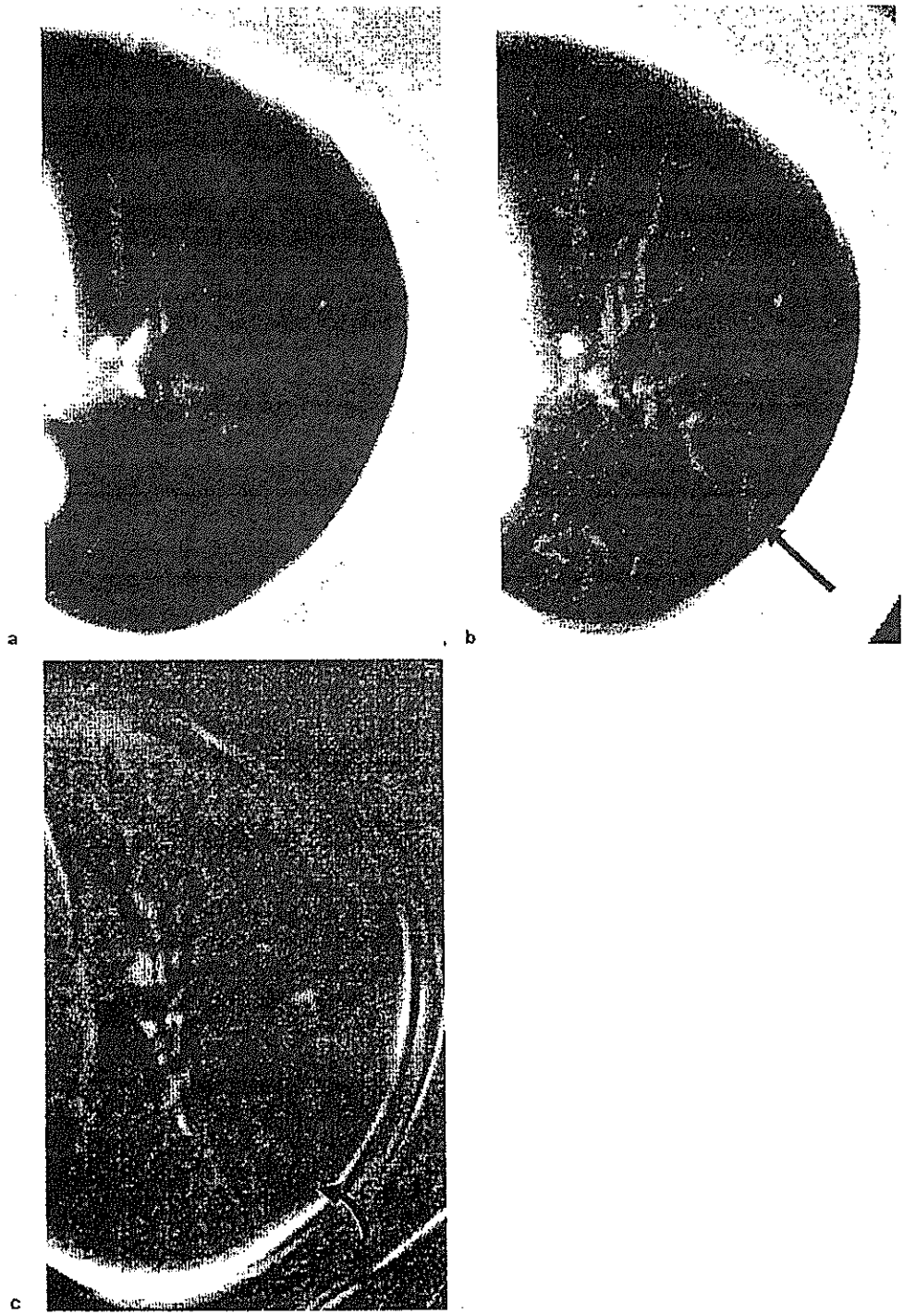


Figure 3. Case 2 **(a, previous CT; b, current CT; c, temporal subtraction image).** A small lung cancer is located in the left peripheral lung field **(b, arrow).** Although temporal subtraction image depicts a small cancer **(c, curved arrow),** misregistration artifacts around the bronchovascular bundle detract from the conspicuity of the lesion.

images without and with aid are shown sequentially. Kobayashi et al (18) and Uozumi et al (10) reported that there were no statistically significant differences in the Az values obtained with the two test methods. In the current study, we used the sequential test because it is more efficient and convenient.

The ROC analysis is a statistically sophisticated way to evaluate the observer performance test. However, the difference in Az values would sometimes be difficult for clinicians to feel its impact for their work. Therefore, some investigators have analyzed the beneficial and detrimental effects by setting the cut-off point on the confidence ratings (10,18–21). The cut-off point is usually set arbitrarily. We believe that this analysis would provide some sense of clinical relevance, if the cut-off point is selected carefully. We selected the cut-off point as 0.5 in the current study. We believe that the cut-off point of 0.5 would be reasonable because the observers were told to use the line for confidence rating uniformly from the point 0 to 1.0, according to their confidence level.

Although we obtained statistically significant differences in observers' performances between without and with temporal subtraction, these results were obtained from a relatively small number of cases (30 cases) with a small number of observers (seven observers). To verify these results, a larger scale of observer test would be warranted.

In conclusion, temporal subtraction may be useful with CT images to improve the diagnostic performance of radiologists in the detection of lung cancer.

REFERENCES

1. Patz EF, Goodman PC, Bepler G. Screening for lung cancer. N Engl J Med 2000; 343:1627–1633.
2. Henschke CI, McCauley DI, Yankelevitz DF, et al. Early Lung Cancer Action Project: A summary of the findings on baseline screening. Oncologist 2001; 6:147–152.
3. 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.
4. White CS, Romney BM, Mason AC, et al. Primary carcinoma of the lung overlooked at CT: Analysis of findings in 14 patients. Radiology 1996; 199:109–115.
5. Gurney JW. Missed lung cancer at CT: imaging findings in nine patients. Radiology 1996; 199:117–122.
6. Kano A, Doi K, MacMahon H, et al. Digital image subtraction of temporally sequential chest images for detection of interval change. Med Phys 1994; 21:453–461.
7. Ishida T, Ashizawa K, Engelmann R, et al. Application of temporal subtraction for detection of interval change in chest radiographs: Improvement of subtraction image using automated initial image matching. J Dig Imaging 1999; 12:77–86.
8. Ishida T, Katsuragawa S, Nakamura K, et al. Iterative image warping technique for temporal subtraction of sequential chest radiographs to detect interval change. Med Phys 1999; 26:1320–1329.
9. Difazio MC, MacMahon H, Xu XW, et al. Effect of time interval difference images on detection accuracy in digital chest radiography. Radiology 1997; 202:447–452.
10. Uozumi T, Nakamura K, Watanabe H, et al. ROC analysis on detection of metastatic pulmonary nodules on digital chest radiographs by use of temporal subtraction. Acad Radiol 2001; 8:871–878.
11. Johkoh T, Kozuka T, Tomiyama N, et al. Temporal subtraction for detection of solitary pulmonary nodules on chest radiographs: Evaluation of a commercially available computer-aided diagnosis system. Radiology 2002; 223:806–811.
12. Lillestrand, RL, Hoyt RR: The design of advanced digital image processing systems. Photogramm Eng 1974; 40:1201–1217.
13. Lillestrand, RL: Techniques for change detection. IEEE Trans C 1972; 21:654–659.
14. Ulstad MS: An algorithm for estimating small scale differences between two digital images. Pattern Recognition 1973; 5:323–333.
15. Metz CE, Herman BA, Shen J-H. Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. Stat Med 1998; 17:1033–1053.
16. Li F, Sone S, Abe H, et al. Missed lung cancers in low-dose helical CT screening obtained from a general population. Radiology 2002; 225: 673–683.
17. Seltzer SE, Judy PF, Adams DF, et al. Spiral CT of the chest: comparison of cine and film-based viewing. Radiology 1995; 197:73–78.
18. Kobayashi T, Xu XW, MacMahon H, et al. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. Radiology 1996; 199:843–848.
19. Kakeda S, Nakamura K, Kamada K, et al. Improved detection of lung nodules by using a temporal subtraction technique. Radiology 2002; 224:145–151.
20. Tsukuda S, Hoshiki A, Katsuragawa S, et al. Detection of lung nodules on digital chest radiographs: Potential usefulness of a new contralateral subtraction technique. Radiology 2002; 223:199–203.
21. Ashizawa K, MacMahon H, Ishida T, et al. Effect of an artificial neural network on radiologists' performance in the differential diagnosis of interstitial lung disease using chest radiographs. AJR Am J Roentgenol 1999; 172:1311–1315.

ORIGINAL ARTICLE

Radiation Medicine: Vol. 22 No. 5, 287–295 p.p., 2004

Pulmonary Lesions Detected in Population-based CT Screening for Lung Cancer: Reliable Findings of Benign Lesions

Tadashi Murakami, Yoshifumi Yasuhara, Shinji Yoshioka, Masahiko Uemura, Teruhito Mochizuki, and Junpei Ikezoe

Purpose: To identify the characteristics of benign pulmonary lesions in order to reduce false-positive rates in screening computed tomography (CT) and in order to reduce frequency of follow-up high-resolution CT (HRCT).

Materials and Methods: We evaluated 238 screening-detected benign lesions and 23 screening-detected lung cancers for 12 characteristics: spiculation, well-defined margin, concave margin, polygonal shape, notch/lobulation, solid component, ground-glass opacity (GGO), air bronchogram, cavity, bubble-like appearance, pleural indentation, and vascular convergence. We also measured the lesion diameters to set a threshold for benign lesions. We tested combinations of these characteristics to differentiate benign lesions from lung cancers.

Results: By using certain combinations of the characteristics that showed statistically significant differences between benign lesions and lung cancers, benign lesions could be extracted without contamination by lung cancer in screening CT, when the combination included solid component as a positive finding. In HRCT, more than 80% of the benign lesions could be extracted without contamination by lung cancer when the combination included GGO as a negative finding.

Conclusion: It seems possible to reduce the frequency of follow-up HRCT to establish a diagnosis of benign lesions using certain combinations of the characteristics of benign nodules.

Key words: lung cancer, screening CT, benign lesions, false positive

Introduction

In mass screenings, computed tomography (CT) is much more sensitive than chest X-ray in detecting lung cancer.[1-7] Screening CT can detect lung cancers at earlier stages than can screening chest X-rays.[1-7] On the other hand, since no effective differentiator between benign nodule and lung cancer has been proposed in screening CT, it detects far more benign nodules (false positives) than screening chest X-rays.[8,9] In fact, the majority of pulmonary nodules detected in CT screening are not lung cancers.[4,9] Therefore, invasive diagnostic procedures for all detected pulmonary lesions cannot be justified. Instead, if screening CT reveals a pulmonary lesion suspected to be lung cancer, diagnostic chest CT with high-resolution CT (HRCT) of the lesion is recommended. If HRCT does not show benign calcification in the lesion, further investigation is recommended, depending on lesion size. Henschke *et al.* reported that, with their protocol, pulmonary nodules detected in CT screening could be managed with little use of invasive diagnostic procedures.[4] To detect growth, that protocol recommends periodic follow-up HRCT or biopsy examination, depending on nodule size.[4] However, to confirm a diagnosis of benign nodule according to the stability of the nodule, follow-up CT is necessary several times for up to 2 years. Since these follow-up CTs are routine practice and are supported by health insurance in Japan, it is preferable and more cost-effective to reduce the false-positive rate in CT screening. If screening CT findings can rule out lung cancer with a reasonable false-negative rate, we can reduce the screening CT false-positive rate. Otherwise, based on the HRCT characteristics of benign nodules,[10] it may be possible to rule out lung cancer at the first

Received January 7, 2004; revision accepted February 21, 2004.

Department of Radiology, Ehime University School of Medicine

Reprint requests to Tadashi Murakami, M.D., Department of Radiology, Ehime University School of Medicine, Shitsukawa, Toon-city, Ehime 791-0295, JAPAN.

A part of this paper was presented at the 61st Annual Meeting of the Japan Radiological Society in 2002.

HRCT examination.

The purpose of the present study was to identify the characteristics of benign pulmonary lesions on screening CT and HRCT, which may reduce the false-positive rate in CT screening or the frequency of diagnostic CT in hospitals. We retrospectively evaluated the results of screening CT and of the first HRCT of cases whose pulmonary lesions were suspicious for lung cancer and for which diagnoses had been established.

Materials and Methods

From December 1999 to March 2003, 16,735 individuals aged over 40 years participated in a mass screening program for lung cancer by mobile spiral CT, which was provided by the Ehime General Health Association. Positive results indicative of lung cancer were found in 2,033 participants (12.1%). Among them, we recruited 195 individuals whose diagnosis had been established and whose screening CT and HRCT were available for this study. Six were excluded from the study because the lesion was acute pneumonia and the shape differed between the screening CT and the HRCT. Finally, the CT images of 189 individuals were analyzed. The mean age was 61 years (range, 42-80 years). In 22 patients (8 men and 14 women), the pulmonary lesion was diagnosed as lung cancer. In the other 167 patients (79 men and 88 women), the pulmonary lesion was diagnosed as benign. The total number of pulmonary lesions was 23 in the lung-cancer patients (one patient had two lung cancers) and 238 in the benign-lesion patients. In all lung-cancer patients, the diagnosis was histopathologically established by surgery. The type of lung cancer was adenocarcinoma in 22 lesions and squamous cell carcinoma in the remaining lesion. Out of 22 adenocarcinomas, the classification of Noguchi was mentioned in seven: type A, 4; type B, 1; type C, 2. The stage of lung cancer was IA in 20 patients, IB in one patient, and IIB in one patient. In the one patient with double-primary cancer (adenocarcinomas), the stage was IB. The diagnoses of a majority of benign lesions were clinically established. Twenty-five lesions showed benign calcification on the HRCT image. During the two-year follow-up period, 209 solid nodules were stable. One lesion was diagnosed as pneumonia because it was improved by the administration of antibiotics. Of the three patients with hamartoma, two were histopathologically diagnosed by biopsy, and the other was diagnosed as having a fat component on the HRCT image (Table 1).

The screening CT was performed with a mobile CT van (Asterion VR, Toshiba Medical Systems, Tokyo). Low-dose spiral CT scanning of the chest was performed with the following parameters: 120 kVp, 25 to 50 mA, 10-mm collimation, 20-mm table speed/rotation. CT images of the entire lung region were obtained in a single, 15- to 20-second breath-hold. The CT images were reconstructed at 10-mm intervals by using a high-frequency algorithm. The image data from the mobile CT were registered in a data-storage system and transferred to a dedicated workstation (RS-252, Konica Co., Tokyo). We interpreted these images at lung window settings (window width 1,600 HU, window level -600 HU) on a monitor with an 8-bit frame memory.

HRCT of the pulmonary lesions was performed with various scanners at the referred hospitals. Although the scan parameters differed among the hospitals, basically both the collimation and reconstruction interval were 1.0-2.0 mm. All image data were reconstructed with a high-frequency algorithm (lung or bone algorithm) and printed on films at lung window settings (window width 1,600 HU, window level -600 HU). We interpreted these images on a view box.

Reading tests were performed separately between screening CT and HRCT. In screening CT, all 238 pulmonary lesions were interpreted because detection of the calcification was impossible. In HRCT, 213 pulmonary lesions without benign calcification were interpreted because a diagnostic workup was terminated whenever a benign calcification was found. In each test, CT images with benign lesions and lung cancers were mixed and interpreted in random order. Three chest radiologists (M.U., S.Y., Y.Y.) interpreted the images and reached a consensus. They did not know the diagnoses of the pulmonary lesions or the patients' clinical information. They also did not know the ratio of benign lesions to lung cancers. They evaluated each pulmonary lesion according to the 12 characteristics that are reported to be useful in differentiating peripheral pulmonary lesions: spiculation, well-defined margin, concave margin, polygonal shape, notch/lobulation, solid component, ground-glass opacity (GGO), air bronchogram, cavity, bubble-like appearance, pleural indentation, and vascular convergence.[10,11] A well-defined margin represents a pulmonary lesion demarcated by a sharp margin from the surrounding lung parenchyma. A concave margin represents a pulmonary lesion with linear or concave sides. A polygonal shape represents a lesion surrounded by a concave margin. The observers rated each pulmonary lesion on the 12 characteristics, assigning a confidence level to each finding as follows: 4, definitely exists; 3, probably exists; 2, probably absent; 1, definitely absent (Figs. 1 and 2).

They recorded the size of each pulmonary lesion by measuring the long axis diameter at the largest section of the pulmonary lesion. In the screening CT, we

Table 1. Summary of materials

	Benign lesions	Lung cancer
Pulmonary lesion	238 lesions in 167 cases	23 lesions in 22 cases
Male/female	79/88	8/14
Age (y.o.)	60.2±0.6	65.2±8.4
Specification	No change: 234	Adenocarcinoma: 22
	Hamartoma: 3	Sq*: 1
	Pneumonia: 1	
		Stage IA: 20
		IB: 1
		IIB: 1

* Sq: squamous cell carcinoma

evaluated the accuracy of this measurement by categorizing the pulmonary lesions into three groups (lesions smaller than or equal to 5 mm,[4] 8 mm,[9] and 10 mm[4]) based on the HRCT results to set a threshold for diagnosing a benign lesion.[4,9] This analysis was performed using lesions without calcification (236 lesions).

To find the characteristics of benign lesions, we compared the frequency of the 12 characteristics and the size differences between the benign lesions and the lung cancers. Then, we tested a combination of the items whose frequencies had differed significantly between the benign lesions and the lung cancers, providing an effective differentiator. We separately analyzed the results in screening CT and HRCT. In this trial, positive findings received scores of 3 or 4 while negative findings were scored 1 or 2. We also performed multivariate analysis to confirm the most effective differentiator between benign lesion and lung cancer.

The Mann-Whitney U-test and Student's t-test were used to compare findings from the screening CT with those of HRCT. A p value of less than 0.05 was regarded as statistically significant.

Results

In screening CT, spiculation ($p<0.05$), GGO ($p<0.01$), air bronchogram ($p<0.01$), bubble-like appearance ($p<0.01$), pleural indentation ($p<0.01$), and vascular convergence ($p<0.01$) were statistically more frequent in lung cancers than in benign lesions. Solid component ($p<0.01$) and polygonal shape ($p<0.05$) were statistically more frequent in benign lesions than in lung cancers. There were no significant differences between benign lesions and lung cancers in notch/lobulation, well-defined margin, concave margin, and cavity. The lung cancers were on average significantly larger than the benign lesions [17.0±9.2 mm (5-42 mm) vs. 6.6±2.8 mm (3-21 mm), respectively] (Fig. 3).

In HRCT, spiculation ($p<0.01$), GGO ($p<0.01$), air bronchogram ($p<0.01$), bubble-like appearance ($p<0.01$), vascular convergence ($p<0.01$), notch/lobulation ($p<0.01$), and pleural indentation ($p<0.01$) were statistically more frequent in lung cancers than in benign lesions. Solid component ($p<0.01$), polygonal shape ($p<0.01$), and well-defined margin ($p<0.01$) were statistically more frequent in benign lesions than in lung cancers. There were no significant differences in concave margin and the presence of cavities between benign lesions and lung cancers. The lung cancers were on average significantly larger than the benign lesions [17.0±9.3 mm (4-35 mm) vs. 6.6±2.5 mm (3-18 mm), respectively] (Fig. 4).

The numbers of pulmonary lesions with below-threshold diameters in HRCT were distributed as follows: smaller or equal to 5 mm, 84 (81 benign lesions, 3 lung cancers) (36%); 8 mm, 191 (185 benign lesions, 6 lung cancers) (81%); 10 mm, 205 (198 benign lesions, 7 lung cancers) (87%). In the screening CT, the numbers of pulmonary lesions with below-threshold diameters were distributed as follows: 5 mm, 82 (80 benign lesions, 2 lung cancers) (35%); 8 mm, 186 (182 benign lesions, 4 lung cancers) (79%); 10 mm, 203 (197 benign lesions, 6 lung cancers) (86%). Between screening CT and HRCT, there was no significant difference in the number of pulmonary lesions under each threshold. However, 24% (20/82) of the lesions below the threshold of 5 mm in the screening CT actually had diameters of greater than 5 mm in HRCT. On the other hand, 97% of the lesions below the threshold of 8 mm and 96% of the lesions below the threshold of 10 mm were identical between the screening CT and HRCT. Thus, we decided to make 8 mm the threshold diameter of benign lesions in the following study.

We used the combination of size threshold and items that had statistically significant differences between lung