Radiologists' Performance for Differentiating Benign from Malignant Lung Nodules on High-Resolution CT Using Computer-Estimated Likelihood of Malignancy

Feng Li[1]
Masahito Aoyama[2]
Junji Shiraishi[1]
Hiroyuki Abe[1]
Qiang Li[1]
Kenji Suzuki[1]
Roger Engelmann[1]
Shusuke Sone[3]
Heber MacMahon[1]
Kunio Doi[1]

OBJECTIVE. The purpose of our study was to evaluate whether a computer-aided diagnosis (CAD) scheme can assist radiologists in distinguishing small benign from malignant lung nodules on high-resolution CT (HRCT).

MATERIALS AND METHODS. We developed an automated computerized scheme for determining the likelihood of malignancy of lung nodules on multiple HRCT slices; the likelihood estimate was obtained from various objective features of the nodules using linear discriminant analysis. The data set used in this observer study consisted of 28 primary lung cancers (6–20 mm) and 28 benign nodules. Cancer cases included nodules with pure ground-glass opacity, mixed ground-glass opacity, and solid opacity. Benign nodules were selected by matching their size and pattern to the malignant nodules. Consecutive region-of-interest images for each nodule on HRCT were displayed for interpretation in stacked mode on a cathode ray tube monitor. The images were presented to 16 radiologists—first without and then with the computer output—who were asked to indicate their confidence level regarding the malignancy of a nodule. Performance was evaluated by receiver operating characteristic (ROC) analysis.

RESULTS. The area under the ROC curve (A_z value) of the CAD scheme alone was 0.831 for distinguishing benign from malignant nodules. The average A_z value for radiologists was improved with the aid of the CAD scheme from 0.785 to 0.853 by a statistically significant level ($p = 0.016$). The radiologists' diagnostic performance with the CAD scheme was more accurate than that of the CAD scheme alone ($p < 0.05$) and also that of radiologists alone.

CONCLUSION. CAD has the potential to improve radiologists' diagnostic accuracy in distinguishing small benign nodules from malignant ones on HRCT.

CT screening has led to early detection of peripheral lung cancer and also detection of a large number of false-positives (i.e., noncalcified benign nodules) [1–5]. The false-positive rate at screening has been reported as 87–93% with low-dose single-detector CT at 10-mm slice thickness [1–3] and 98–99% with single-detector CT or MDCT at 5-mm slice thickness [4, 5]. Also, simultaneous or additional diagnostic high-resolution CT (HRCT) is needed for the distinction between malignant and benign lung nodules detected as suspicious or indeterminate lesions on screening CT [1–5]. This high false-positive rate because of benign nodules is likely to reduce the benefit of CT screening for early detection of lung cancer [6]. Therefore, it is important to differentiate benign from malignant nodules to reduce the number of false-positive findings on screening CT and to reduce follow-up examinations for diagnostic HRCT.

We developed an automated computerized scheme [7] for determination of the likelihood measure of malignancy by using various objective features of the nodules in our a database of thick-section low-dose CT; one or two slices were used for image analysis of each nodule. The low-dose CT database consisted of 489 nodules obtained from a mass screening for lung cancer in Nagano, Japan [2]. All of these nodules were considered as suspicious or indeterminate lesions when detected by radiologists on low-dose CT screening. With the use of receiver operating characteristic (ROC) analysis, our computerized scheme achieved an area under the ROC curve (A_z value) of 0.846 for distinction between 76 malignant and 413 benign lung nodules.

Recently, we further developed another computerized scheme for distinction be-

Received November 10, 2003; accepted after revision January 30, 2004.

Supported in part by USPHS grant CA62625.

H. MacMahon and K. Doi are shareholders of R2 Technology, Inc., Los Altos, CA. K. Doi is a shareholder of Deus Technology, Inc., Rockville, MD.

[1]Department of Radiology, Kurt Rossmann Laboratories for Radiologic Image Research, MC-2026, The University of Chicago, 5841 S Maryland Ave., Chicago, IL 60637. Address correspondence to F. Li (fli@kurt.bsd.uchicago.edu).

[2]Department of Intelligent Systems, Faculty of Information Sciences, Hiroshima City University, Hiroshima 731-3194, Japan.

[3]Azumi General Hospital, Ikeda, Nagano 399-8695, Japan.

AJR 2004;183:1209–1215

0361–803X/04/1835–1209

tween malignant and benign lesions derived from multiple slices of HRCT (1-mm collimation) based on 2D and 3D volume data. The HRCT database consisted of 244 small noncalcified (3–20 mm) nodules obtained as part of follow-up diagnostic work for suspicious or indeterminate lesions detected on low-dose CT in the same screening program.

In the present study, we assessed observer performance using ROC analysis to evaluate the effectiveness of our computer-aided diagnosis (CAD) scheme to assist radiologists in distinguishing small benign from malignant lung nodules in various patterns at HRCT. The malignant lung cancers included nodules with pure ground-glass opacity, mixed ground-glass opacity, and solid opacity; the benign nodules were selected by matching their size and pattern to the cancers on HRCT in this observer study.

Materials and Methods

Our institutional review board approved the use of this database and the participation of radiologists in this observer performance study. Informed consent for use of cases was waived. Informed consent for the observer performance study was obtained from all observers.

Database

The diagnostic HRCT database used in this study consisted of 59 patients (27 men, 32 women; mean age, 64.6 years) with 61 malignant nodules and 169 patients (99 men, 70 women, mean age 61.6 years) with 183 benign nodules. The database was obtained as part of an annual 3-year low-dose CT screening for lung cancer from 17,892 examinations on 7,847 individuals in Nagano, Japan [2]. HRCT scans were obtained on a helical scanner (HiSpeed Advantage, GE Healthcare) with a standard tube current (200 mA) to cover the entire nodule lesion, 1-mm collimation, and a bone reconstruction algorithm with a 0.5-mm interval.

Two features concerning the size and pattern type of the pulmonary nodules on HRCT were subjectively determined by radiologists for the purpose of grouping nodules in our database. The mean size (average length and width) was recorded by one radiologist. The three types of patterns of these nodules—pure ground-glass opacity, mixed ground-glass opacity, and solid opacity—were viewed independently and grouped by three radiologists without knowledge of the final diagnosis, and a consensus was reached through discussion. Nodules with benign-pattern calcifications (diffuse, central, popcorn, and laminar, or concentric calcification) were excluded. The range of nodule sizes for the 61 malignant and 183 benign nodules was 6–19 mm (mean, 12 mm) and 3–20 mm (mean, 7 mm), respectively. Among the 61 malignant nodules, there were 18 nodules with pure ground-glass opacity, 28 with mixed ground-glass opacity, and 15 with solid opacity, whereas 183 benign nodules included 12 with pure ground-glass opacity, 30 with mixed ground-glass opacity, and 141 with solid opacity.

All malignant nodules were primary lung cancers confirmed by surgery, including 49 well-differentiated adenocarcinomas, eight other adenocarcinomas, two squamous cell carcinomas, and two localized small cell carcinomas. Among the 183 benign nodules, nine (four cases of nodular fibrosis; and one case each of inflammatory granuloma, cryptococcoma, focal organizing pneumonia, inflammatory pseudotumor, and sclerosing hemangioma) were confirmed by surgery, 51 had resolved at follow-up examination, and 123 showed no change for 2 or more years.

CAD

With our CAD scheme, the nodules were segmented automatically using a dynamic programming technique [7]. Forty-one and 15 image features based on 2D and 3D volume data, respectively, were determined from quantitative analysis of the nodule outline and pixel values. Linear discriminant analysis was used to distinguish benign from malignant nodules. The performance of this CAD scheme was evaluated on the basis of a leave-one-out testing method by use of 61 malignant and 183 benign lung nodules. For the input of the linear discriminant analysis, we selected many combinations from 56 features and two clinical parameters (patient age and sex). The following features were used in this study: effective diameter, contrast of the segmented nodule on the HRCT image, overlap measures of two gray-level histograms for the inside and outside regions of the segmented nodule on the HRCT image, overlap measures of two gray-level histograms for the inside and outside regions of the segmented nodule on the edge-gradient image, radial gradient index for the inside region of the segmented nodule on the HRCT image, peak value of the histogram for the inside region of the segmented nodule on the edge-gradient image, pixel value at the peak of the histogram for the inside region of the segmented nodule on the edge-gradient image, and pixel value at the peak of the histogram for the inside region of the segmented nodule on the HRCT image.

Our computerized classification method outputs a percentage, from 1% to 99%, that indicates the likelihood of malignancy. The performance of the classification scheme yielded an A_z value of 0.937 (0.919 for nodules with pure ground-glass opacity, 0.852 for nodules with mixed ground-glass opacity, and 0.957 for solid nodules) for distinction between 61 malignant and 183 benign lung nodules.

Observer Study

The data used in this observer study consisted of 28 malignant nodules that were randomly selected from the 61 primary lung cancers and 28 benign nodules that were selected from the 183 benign nodules by matching in size and pattern to the cancers. For both malignant and benign lesions, nine nodules ranged from 6 to 10 mm and 19 nodules ranged from 11 to 20 mm. The patterns involved were eight nodules with pure ground-glass opacity, 12 with mixed ground-glass opacity, and eight with solid opacity. Examples of cases used for this observer study are shown in Figure 1.

Sixteen radiologists participated in this observer study. The 16 radiologists, seven chest radiologists and nine other radiologists, have a mean of 14 years of experience (range, 7–26 years). Consecutive region-of-interest HRCT images for each nodule were displayed for interpretation using the cine mode on a cathode ray tube monitor ($1,280 \times 1,024$ resolution). The window settings were initially at a width of 1,500 H and a level of –550 H, but the settings could be adjusted by the observer. In addition, zooming capability was provided. Two clinical parameters (patient age and sex) were displayed to the observer on the monitor.

The observers were told that the purpose of this observer study was to assist radiologists in distinguishing benign from malignant lesions on HRCT by using a CAD scheme. The instructions for the observers were an explanation of the role of CAD output as a second opinion. The observers were told that 28 malignant lesions (6–10 mm, nine cases; 11–20 mm, 19 cases; pure ground-glass opacity, eight cases; mixed ground-glass opacity, 12 cases; and solid opacity, eight cases) and 28 benign lesions (matched in size and pattern to the malignant lesions) were included in this study and that the sensitivity and specificity of our CAD scheme, using a threshold of 0.50 (50%) likelihood of malignancy, are 80% and 75%, respectively.

The observers were instructed to click on a bar (left, benignancy; right, malignancy) on the screen using a mouse to indicate confidence level regarding the malignancy (or benignancy) of a lesion first without and then with computer output, and after indicating your confidence (without and with CAD), click on one of the four following clinical actions: return to annual screening; follow-up in 6 months; follow-up in 3 months; or biopsy or surgery.

For a training session before the test, we provided five 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. The review time was not limited. The average review time was 46 min (range, 28–100 min).

Data Analysis

The confidence level ratings from each observer were analyzed using receiver operating characteristic (ROC) methodology, and a quasimaximum likelihood estimation of the binormal distribution was fitted to the radiologists' confidence ratings [8]. The statistical significance of the difference in A_z values between observer interpretations without and with the CAD scheme was tested using the Dorfman-Berbaum-Metz method [9]; this method included both observer variation and case sample variation by means of an analysis-of-variance approach. The sta-

tistical significance of the difference in A_z values between the computer outputs and observer interpretations (without and with the CAD scheme) was tested by means of confidence interval method by taking into account observer variation alone [10]. The effect of the computer output on the rating scores and also the change in scores that were due to the use of the CAD scheme were analyzed. The distributions of the radiologists' ratings and of the computer outputs were compared for the malignant and benign nodules.

The statistical significance of the difference in clinical actions between the beneficial and detrimental effect of the CAD scheme for each of the malignant and benign nodules was estimated using the Student's paired *t* test for 16 radiologists.

Results

For the cases selected for this observer study, the A_z value of the CAD scheme alone was 0.831 for distinguishing 28 malignant and 28 benign nodules (0.910 for nodules with pure ground-glass opacity, 0.814 for nodules with mixed ground-glass opacity, and 0.783

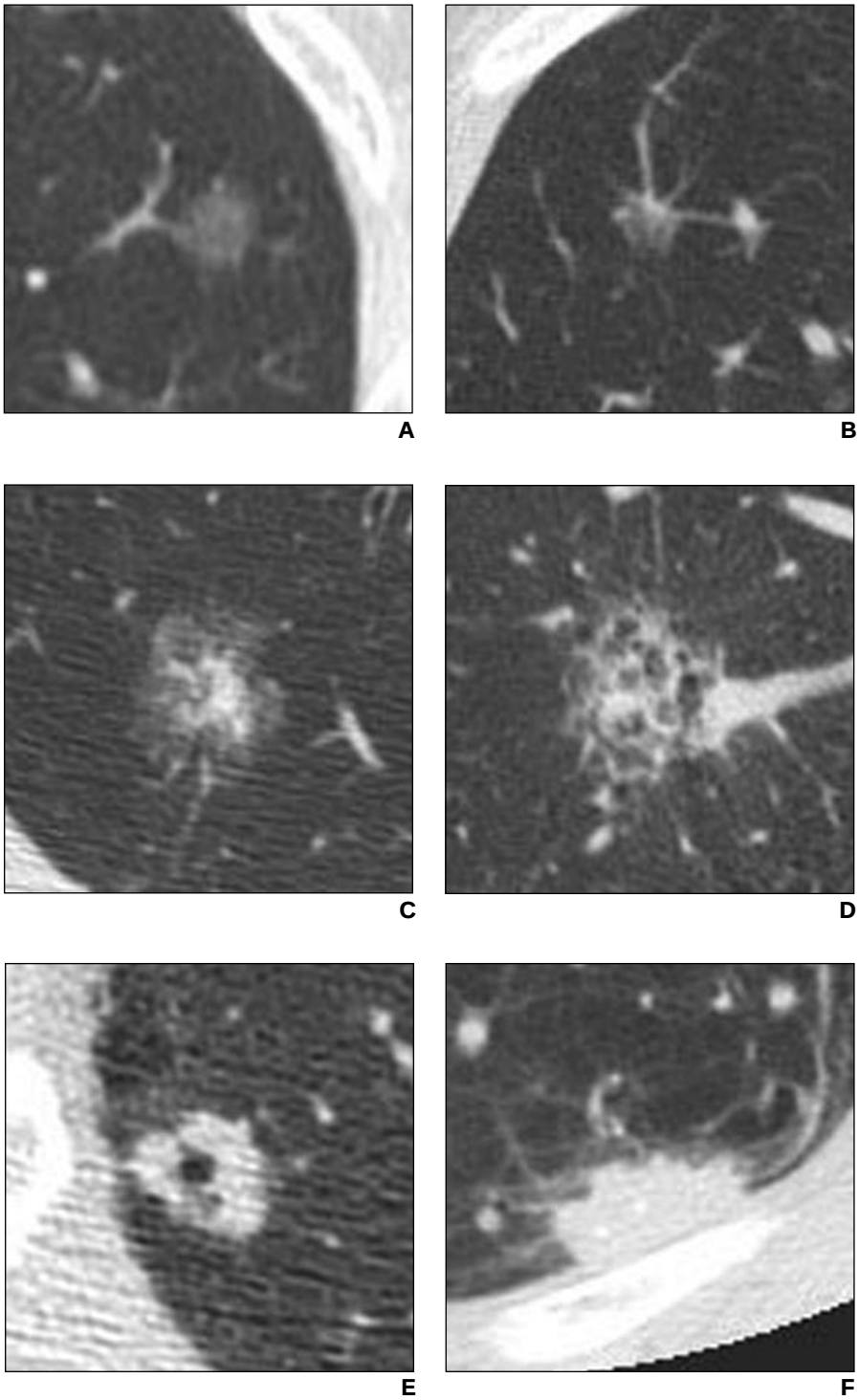

Fig. 1.—Radiologists' average ratings without and with computer output for six cases used in observer study. Note that difference in likelihood of malignancy between computer output and initial radiologists' ratings was not large in cases shown here. Radiologists' interpretation with computer-aided diagnosis (CAD) scheme was, in general, more accurate than radiologists without CAD scheme in most malignant and benign nodules.
A, High-resolution CT (HRCT) scan of 55-year-old woman with lung cancer shows pure ground-glass opacity. Computer output was 0.66; radiologists' ratings without CAD, 0.64; and radiologists' ratings with CAD, 0.71.
B, HRCT scan of 57-year-old woman with benign nodule shows pure ground-glass opacity. Computer output was 0.24; radiologists' ratings without CAD, 0.32; and radiologists' ratings with CAD, 0.27.
C, HRCT scan of 73-year-old man with lung cancer shows mixed ground-glass opacity. Computer output was 0.90; radiologists' ratings without CAD, 0.75; and radiologists' ratings with CAD, 0.85.
D, HRCT scan of 79-year-old man with benign nodule shows mixed ground-glass opacity. Computer output was 0.57; radiologists' ratings without CAD, 0.48; and radiologists' ratings with CAD, 0.56.
E, HRCT scan of 57-year-old man with lung cancer shows solid opacity. Computer output was 0.78; radiologists' ratings without CAD, 0.66; and radiologists' ratings with CAD, 0.76.
F, HRCT scan of 68-year-old man with benign nodule shows solid opacity. Computer output was 0.36; radiologists' ratings without CAD, 0.37; and radiologists' ratings with CAD, 0.36.

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

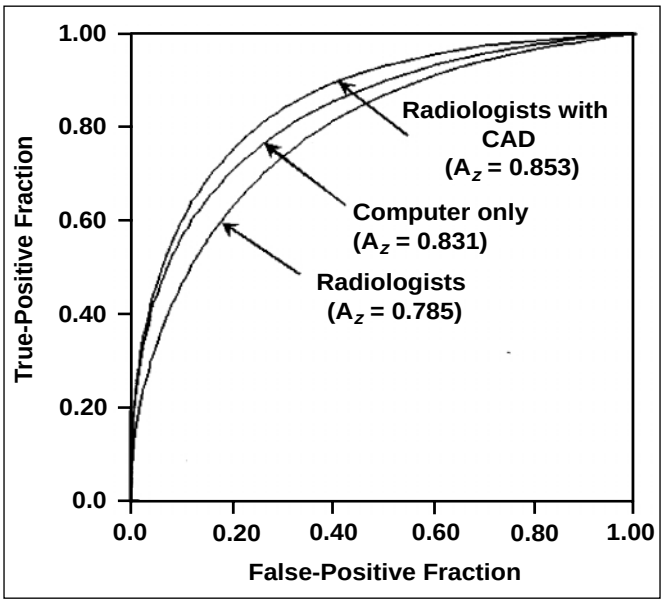

Fig. 2.—Graph shows receiver operating characteristic (ROC) curves for performance of computer alone and average performance of 16 radiologists without and with computer-aided diagnosis (CAD) scheme. Note that difference was statistically significant between radiologists without and with CAD scheme ($p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

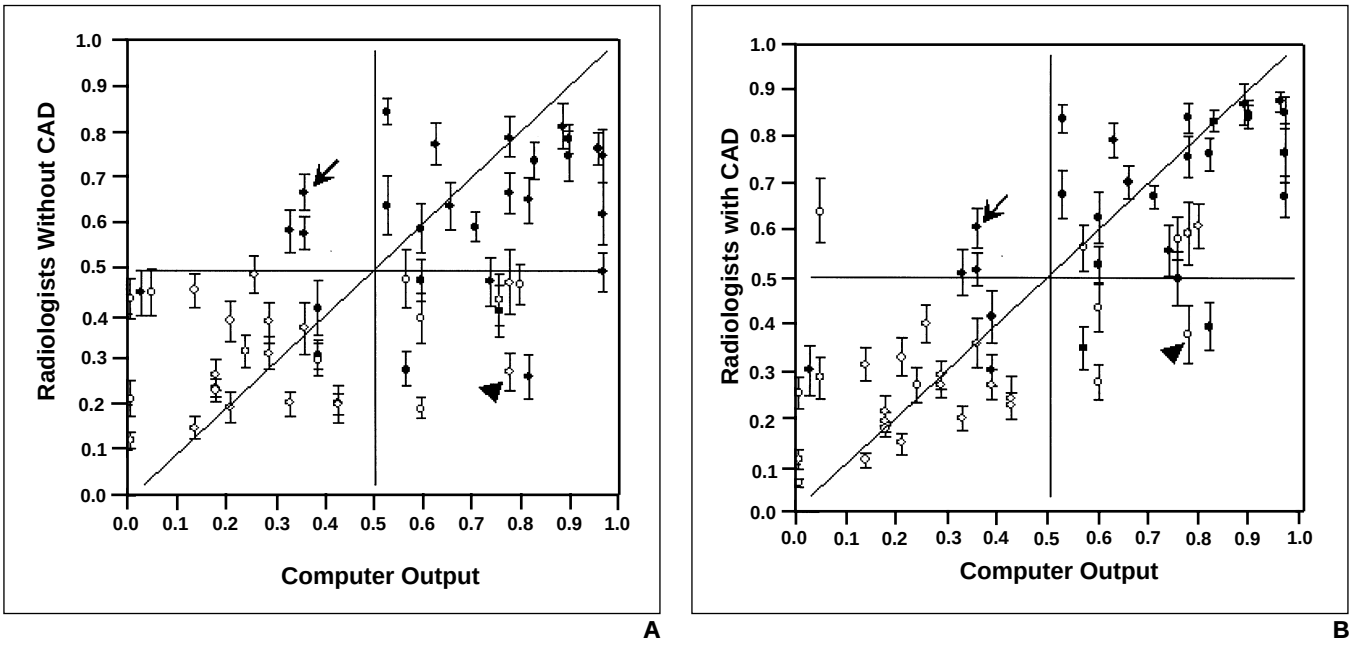

Fig. 3.—Graphs show correlation between computer output and average radiologists' ratings without and with computer-aided diagnosis (CAD) scheme for indicating likelihood of malignancy for lung nodules. ● = average radiologists' ratings for malignant nodules, ○ = average radiologists' ratings for benign 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-

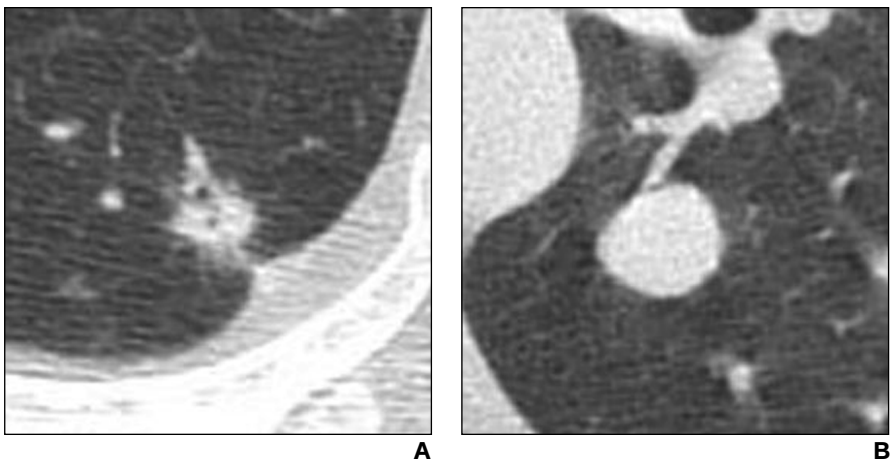

Fig. 4.—High-resolution CT (HRCT) scans show one malignant case and one benign case. Note that radiologists' interpretations without computer-aided diagnosis (CAD) scheme were correct in these cases, whereas likelihoods of malignancy based on computer outputs only were obviously incorrect; even with 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.

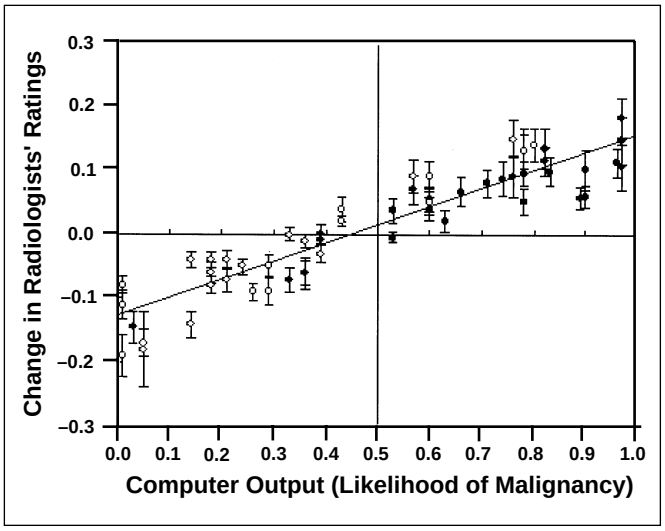

Fig. 5.—Graph shows correlation (r = 0.925) between likelihood of malignancy and average change in confidence level (rating scores) for each nodule by 16 radiologists. Malignant and benign nodules are marked by black circles and white circles, respectively. ● = average change in confidence level for 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, Vyborny 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