Published in final edited form as:

Acad Radiol. 2010 March ; 17(3): 323–332. doi:10.1016/j.acra.2009.10.016.

Computer-Aided Diagnosis of Lung Nodules on CT Scans: ROC Study of Its Effect on Radiologists' Performance

Ted Way, PhD, Heang-Ping Chan, PhD, Lubomir Hadjiiski, PhD, Berkman Sahiner, PhD, Aamer Chughtai, MD, Thomas K. Song, MD, Chad Poopat, MD, Jadranka Stojanovska, MD, Luba Frank, MD, Anil Attili, MD, Naama Bogot, MD, Philip N. Cascade, MD, and Ella A. Kazerooni, MD

Department of Radiology, University of Michigan, Med-Inn Building C477, 1500 E Medical Center Drive, Ann Arbor, MI 48109-5842 (T.W., H.-P.C., L.H., B.S., A.C., J.S., L.F., A.A., N.B., P.N.C., E.A.K.); the Department of Radiology, Henry Ford Hospital, Detroit, MI (T.K.S., C.P.).

Abstract

Rationale and Objectives—The aim of this study was to evaluate the effect of computer-aided diagnosis (CAD) on radiologists' estimates of the likelihood of malignancy of lung nodules on computed tomographic (CT) imaging.

Methods and Materials—A total of 256 lung nodules (124 malignant, 132 benign) were retrospectively collected from the thoracic CT scans of 152 patients. An automated CAD system was developed to characterize and provide malignancy ratings for lung nodules on CT volumetric images. An observer study was conducted using receiver-operating characteristic analysis to evaluate the effect of CAD on radiologists' characterization of lung nodules. Six fellowship-trained thoracic radiologists served as readers. The readers rated the likelihood of malignancy on a scale of 0% to 100% and recommended appropriate action first without CAD and then with CAD. The observer ratings were analyzed using the Dorfman-Berbaum-Metz multireader, multicase method.

Results—The CAD system achieved a test area under the receiver-operating characteristic curve (A_z) of 0.857 ± 0.023 using the perimeter, two nodule radii measures, two texture features, and two gradient field features. All six radiologists obtained improved performance with CAD. The average A_z of the radiologists improved significantly ($P < .01$) from 0.833 (range, 0.817–0.847) to 0.853 (range, 0.834–0.887).

Conclusion—CAD has the potential to increase radiologists' accuracy in assessing the likelihood of malignancy of lung nodules on CT imaging.

Keywords

Computer-aided diagnosis; pulmonary nodule; observer study; computed tomography

The use of computed tomographic (CT) imaging for lung cancer screening is an area of active research. CT imaging has been shown to be more sensitive for lung nodule detection than chest x-ray, especially for smaller nodules (1–4). Henschke et al (5) reported a 92% survival rate among patients who underwent surgical resection for detected stage I lung cancers. Sobue et al (6) reported an almost 100% 5-year survival rate for patients with nodules < 9 mm in size. These data suggest the benefits of earlier intervention with early

detection. It is expected that the National Lung Screening Trial (7), which is a randomized, controlled study of >50,000 enrolled patients, will provide more definitive results as to whether early detection with CT imaging compared to CXR will lead to reduced patient mortality.

The higher sensitivity of CT imaging results in an increase in the number of nodules detected and thus an increase in the nodules that need to be followed up and managed. This may require expensive diagnostic tests such as follow-up CT scans and biopsies. Multiple-detector row CT technology has resulted in thinner slices and higher resolution. However, the large number of images that radiologists must interpret greatly increases their workload. Despite higher quality images, Swensen et al (8) reported that as many as 50% of nodules resected at surgery are benign, signifying the difficulty radiologists have in determining whether a lung nodule is malignant or not on the basis of CT and other clinical information. This emphasizes the importance of developing methods that can assist radiologists in characterizing nodules.

Computer-aided detection and diagnosis software is being developed to address these issues. Computer-aided detection has been shown to increase the sensitivity of lung nodule detection (9–14). Computer-aided diagnosis (CAD) methods for the classification of lung nodules as malignant or benign have been reported by a number of investigators, with the area under the receiver-operating characteristic (ROC) curve, A_z , ranging from 0.79 to 0.92 (15–23). A few observer studies have been performed to evaluate the effects of CAD on radiologists' assessments of the malignancy of lung nodules on CT images. Matsuki et al (24) performed a study with four radiologists, four fellows, and four residents reading a data set of 25 malignant and 25 benign nodules. They found that the average A_z of the entire group increased significantly from 0.831 to 0.959 ($P < .001$) with CAD. The average A_z of each group of observers also improved significantly, and the differences in performance among the three groups were reduced. Shah et al (25) conducted a study with eight radiologists reading 28 nodules (15 malignant, 13 benign) and obtained a significant improvement in the average A_z from 0.75 to 0.81 ($P = .018$) when a computer aid was used. Li et al (21) found that the average A_z for 16 radiologists significantly increased from 0.785 to 0.853 ($P = .016$). In addition, they observed that CAD had a beneficial effect on 68% of their changed recommendations for a data set of 28 malignant and 28 benign nodules (26). Awai et al (27) reported a significant ($P = .021$) improvement for 19 observers from an average A_z of 0.843 ± 0.097 to 0.924 ± 0.043 for 18 malignant and 15 benign nodules. A subgroup analysis showed that the nine radiology residents improved significantly as a group, but the improvement of 10 board-certified radiologists did not achieve statistical significance.

Although the previous studies demonstrated a trend of improvement in radiologists' classification accuracy with CAD, the data sets in those studies were small. In this study, we collected a relatively large data set of 256 (132 benign, 124 malignant) nodules including both primary and metastatic lung cancers and a range of CT parameters to evaluate the effect of our CAD system on radiologists' estimates of the malignancy of lung nodules. The aim was to reveal the effect of CAD in an environment with a heterogeneous case mix in comparison to that of a homogeneous data set.

METHODS AND MATERIALS

Collection of CT Studies

We retrospectively collected CT scans from the patient archive in the radiology department at our institution. The CT studies were acquired in our clinic with a variety of GE scanners (GE Healthcare, Waukesha, WI), including the Genesis HiSpeed scanners and the GE

LightSpeed series scanner models Plus, Power, Pro 16, QX/i, Ultra, and Light-Speed 16. The pixel size ranged from 0.448 to 0.859 mm (with corresponding fields of view of 25–44 cm). Slice thickness averaged 2.3 ± 1.44 mm (range, 1–7.5 mm), and the slice interval averaged 2.0 ± 1.6 mm (range, 0.6–7.5 mm). The average values for the scanning parameters were 120 kVp for tube voltage (range, 120–140 kVp), and 214 ± 141 mAs (range, 40–570 mAs) for tube current–time product. Our study was approved by the institutional review board and in compliance with the Health Insurance Portability and Accountability Act.

Nodule Selection

For each patient scan, an expert thoracic radiologist marked the locations of nodules by placing a box encompassing each nodule using a graphical user interface (GUI) developed in our laboratory. This radiologist did not participate as an observer. The nodule inclusion criteria for this study were (1) diameter > 3 mm as measured by the radiologist, (2) appearance of the nodule on at least three consecutive slices, and (3) proven diagnosis through biopsy, other known metastatic disease, or 2-year follow up.

We collected 256 nodules from the CT scans of 152 patients. There were 132 benign and 124 malignant nodules. Because of the invasiveness of the lung biopsy procedure, clinicians generally do not perform a biopsy for every lung nodule in clinical practice. The nonbiopsied nodules were determined to be benign, primary cancer, or metastatic cancer by clinicians using all available diagnostic information during the patients' clinical care. The original diagnosis in the clinical reports and any additional follow-up information available by the time of our data collection were used as reference to determine whether a nonbiopsied nodule in our data set should be labeled benign, primary, or metastatic cancer. Of the 124 malignant nodules, 64 were established by biopsy, and 60 were determined to be malignant as described above. Seventy-two of the malignant nodules were primary and 52 were metastatic cancers. Of the 132 benign nodules, 15 were biopsy proven and 117 were determined to be benign by 2-year follow-up stability on CT imaging. There were 218 solid nodules, 17 nodules with ground-glass opacity, and 21 mixed attenuation types. Of the 256 nodules, 53 were juxtapleural and 19 were juxtavascular. A distribution of the longest diameters of the nodules measured by radiologists is shown in Figure 1. The nodules had an average longest diameter of 11.7 ± 7.7 mm (range, 3.0–37.5 mm). Eight of the nodules in our data set had longest diameters > 30 but < 38 mm. Nodules are generally defined as < 30 mm in diameter, but we included these masses in our data set because their edges and surrounding texture may contribute to the training of the CAD system. Although they seem highly suspicious, one was proven benign by biopsy. Test results with and without these eight masses were compared.

This study was designed to simulate the situation in which a radiologist reads the first scan of a patient and makes a recommendation on whether a follow-up scan or other immediate action needs to be taken, so that no serial CT scan, and thus interval change information, was available for either the radiologist or the CAD system.

CAD System

Our CAD system is summarized as follows; further details can be found in the literature (15,28). First, a volume of interest (VOI) containing the nodule was extracted on the basis of the box placed by the expert radiologist. Because our CAD system was designed to classify whether a nodule was malignant or benign, the input to the system was assumed to be a VOI that contained a lung nodule. The system segmented the nodule from the surrounding background tissue within the VOI using three-dimensional clustering and a three-dimensional active-contour method. Morphologic features including volume, largest

perimeter, and statistics based on the CT values (Hounsfield units) inside the nodule were then extracted from the segmented nodule. To quantify tissue texture around the nodule, the rubber band straightening transform (29) was used to convert a 15-voxel-wide band surrounding the nodule on each slice into a rectangular image. After performing Sobel filtering on the rubber band straightening transformed images, run-length statistics features (30,31) were extracted. In addition, gradient field features (28,32) were extracted from the gradient magnitude value at each voxel. The statistics of the gradient magnitudes of all surface voxels and along the rays tracing from the nodule centroid to the surface voxels were used to describe the smoothness of the nodule surface.

We used a “two-loop” leave-one-case-out resampling method to train and test the CAD system using the n available cases (28). In each cycle of the outer leave-one-case-out loop, we reserved one case, including all nodules from this case, as the independent test case. The remaining $n - 1$ cases were used to train the classifier in a process that included feature selection and classifier weight determination. A subset of most effective features was selected by stepwise feature selection. An “inner” leave-one-case-out scheme was performed within the $n - 1$ training cases to determine the best thresholds for stepwise feature selection. These thresholds were F_{in} and F_{out} for deciding whether a feature should be included or removed from the feature space, respectively, and the tol threshold for setting the tolerance on the correlation of the selected features. In each cycle of this inner leave-one-case-out scheme for feature selection, $n - 2$ cases were available for training, while one case was left out as the test case. The best set of F_{in} , F_{out} , and tol thresholds was searched by simplex optimization using the test A_z of the $n - 1$ left-out cases from the inner loop as a guide. After the feature selection thresholds were determined, a set of features was selected from the $n - 1$ cases, and a linear discriminant analysis classifier with proper weights for the features was built. This classifier was then applied to all nodules of the original independent left-out case and a test score for each nodule was obtained. This procedure was cycled through the n cases of the entire data set in the outer loop, so that each case was left out in turn, resulting in test scores for all the nodules in the data set.

A histogram with 10 bins was generated from the test scores of the entire data set. Each bin was further separated into benign and malignant classes. For each class, a Gaussian curve was fitted (SigmaPlot 9.0; SysStat, San Jose, CA). Both curves were normalized so that the area under each curve was unity (Fig 2). The two fitted Gaussian curves represented the probability density functions of the malignancy ratings for test lung nodules estimated by the CAD system. The original bin value was mapped linearly as a relative malignancy rating on a scale of 1 to 10. The two fitted class distributions were shown to the radiologist as a reference together with the malignancy rating during the reading with CAD for a given nodule.

Observer Study

We conducted an ROC study with a sequential reading method in which the radiologist was asked to estimate the likelihood of malignancy (LM) of a nodule, view the malignancy rating by the CAD system, and then modify his or her LM estimate if desired. The sequential reading method emulated the use of CAD as a second opinion, in which the radiologist first made his or her own judgment without CAD and then made a refined decision after taking the malignancy rating of the CAD system into consideration. Six fellowship-trained thoracic radiologists who had an average of 3.2 postfellowship years (range, 1–8 years) of experience interpreting thoracic CT images participated as observers.

The reading order of the nodules was “randomized” and counterbalanced for each reader such that, on average, no nodule would be read more often in a certain order (eg, first or last) in the reading sessions than the other nodules. Different nodules from the same case were

separated by a number of nodules from other cases, so that no two nodules from the same patient were presented consecutively. The randomization and counterbalanced design were intended to minimize the effects of fatigue, learning, memorization, and nodule correlation on the results of observer performance (33).

We developed a GUI (Fig 3a) to display the CT scan and record the observer ratings in this study. One nodule was read at a time, but the entire CT scan was loaded and the slice interval was displayed. To ensure that the observer read the correct nodule, the nodule of interest was marked by a box, the location of which was previously determined by an expert thoracic radiologist during nodule selection, who did not participate as a reader. The slice in which the nodule appeared to be the largest was shown first, but the observer was asked to scroll through all slices of the nodule to evaluate its characteristics, as in clinical practice. The box size was large enough so that the margin characteristics of the nodule were not obscured by the box, as shown in the example in Figure 3a. The overlaying box could also be turned off by a click if the observer preferred. The observer was free to scroll through the available slices from the scan, but he or she was instructed to focus on the visual characteristics of the nodule of interest. No clinical or demographic information about the patient was provided. Readers were free to adjust the brightness and contrast of the image, and a zoom function was available. A 3×3 mm box displayed in the upper left corner of the image served as a size reference. A rendered volume for each nodule was available should the observer choose to look at its surface characteristics.

Each radiologist was asked to rate the LM of the marked nodule on a scale of 0% to 100% and provide the recommended action (no action, CT follow-up, or immediate action, such as biopsy, positron emission tomography, or surgery). In addition, the radiologist provided feature descriptors for the nodule, including the presence of cavitation, calcification, nodule edge (smooth, lobulated, or spiculated or irregular), and attenuation type (solid, ground-glass opacity, or mixed).

The GUI prevented the reader from viewing the rating of the CAD system until the assessments listed above were completed. The classifier result was presented as an integer rating on a scale of 1 to 10 (Fig 3b), as described in the previous section. The probability density distributions of malignant and benign classes as estimated by the CAD system were shown on the GUI to provide a reference for the observer. After viewing the CAD system rating, the radiologist had the option of adjusting the LM estimate of the nodule and the recommended action.

Each observer underwent a training session with nodules not part of the data set to become familiar with the GUI and the experimental process before the actual reading session would start. We instructed the observers to use the entire range of the rating scale and to interpret the CAD system rating by reference to the two-class distributions of the classifier. The radiologists were informed of the total number of nodules and the number of patients. They were not told the proportion of malignant and benign nodules, only that the prevalence of malignant nodules was enriched compared to what they would see in clinical practice. No time limit was imposed to assess each nodule.

Statistical Analysis

We analyzed the radiologists' malignancy ratings using ROC methodology. The classification accuracy was quantified by A_z , which was estimated using the Dorfman-Berbaum-Metz method for the analysis of multireader multicase data (34). The Dorfman-Berbaum-Metz method uses the maximum likelihood estimation of the binormal distributions to fit the observer rating data and provides an estimate of the statistical significance of the difference in the two conditions, without and with CAD, taking into

account the multireader multicase readings. The slope and intercept parameters of the individual observers' ROC curves were averaged, and these average parameters were used to derive an averaged ROC curve (33). We also calculated the partial A_z , denoted $A_z^{0.9}$, which is the area under the ROC curve above a true-positive fraction value of 0.9. A larger value of $A_z^{0.9}$ indicates higher specificity in the high-sensitivity region (35,36). The significance in the difference between the observer's $A_z^{0.9}$ values without and with CAD was estimated using Student's two-tailed paired t test.

Because primary cancers and metastatic cancers have somewhat different characteristics and may be distinguished from benign nodules in different ways, we separately analyzed the classification accuracy for two subsets of lung cancers; one subset contained only the primary cancers and the benign nodules, and the other subset contained the metastatic cancers and the benign nodules to study whether the performance of radiologists might be different for nodules with different characteristics. In addition, we analyzed the performance when the eight masses in the data set > 30 mm in diameter were excluded to evaluate the classification of lesions that were considered to be nodules (< 30 mm) by radiologists.

Because some of the nodules in our data set were obtained from the same CT scan, we used the Obuchowski method for the analysis of A_z values for clustered data (37), which was generalized to multireader multimodality studies by Lee and Rosner (38), to account for the possible correlations between the rating data of nodules from the same scan. The P values estimated by the Obuchowski method were reported when the areas under the ROC curves from the without-CAD and with-CAD reading conditions were compared. Similar corrections were performed in our previous ROC analysis (39).

RESULTS

The CAD system achieved a two-loop leave-one-case-out test A_z of 0.857 ± 0.023 for the 256 nodules and a partial A_z , $A_z^{0.9}$, of 0.476. The selected features were very consistent, with only a slight variation among the 152 (the total number of cases) cycles in the leave-one-case-out process. An average of 6.62 features were selected, and the most consistently selected features were perimeter, the skewness of the gradient magnitude values of the surface voxels, a profile feature, the variance of radii segments, the skewness of all radii segments, and two run-length statistics texture features. In the 152 cycles, these features were each selected between 150 and 152 times, except for one texture feature that was selected in 86 cycles. This shows that the features were consistently good features, and minor variations in the training set did not drastically change the set of features that would be selected. Detailed analysis of the CAD system performance was reported elsewhere (28).

For the radiologists' performance, the average A_z without CAD was 0.833 (range, 0.817–0.847), and it improved significantly to 0.853 (range, 0.834–0.877) with CAD ($P < .01$). In addition, $A_z^{0.9}$ improved significantly from 0.390 to 0.456 ($P = .043$). All radiologists showed improvement in terms of A_z . The A_z values for the radiologists are shown in Table 1. The average ROC curves for the radiologists without and with CAD, in addition to the ROC curve of the computer classifier, are compared in Figure 4.

Of the 256 nodules, the radiologists modified their LM estimates after the use of CAD an average of 126.0 ± 46.8 times (range, 57–192 times). We define a "correct" LM change when a radiologist increased the LM estimate for a malignant nodule or reduced the LM estimate for a benign nodule with CAD, and vice versa for an "incorrect" change. The radiologists made correct LM changes an average of 95.0 ± 34.0 times (range, 37–126 times) out of the 126 average changes, modifying their estimates by an average of 10.2 ± 2.8

points (range, 6.6–113.4 points). The radiologists made incorrect LM changes an average of 31.0 ± 18.2 times (range, 16–66 times), changing their estimates by an average of 10.9 ± 3.6 points (range, 6.7–16.2 points). These changes are summarized in Table 2.

At present, there is no guideline on the threshold of the LM at which a nodule should be considered malignant or benign. As an example, we chose 50% as an arbitrary cutoff to demonstrate the number of times the radiologists changed their decisions on nodules being malignant or benign in our study. For malignant nodules, the changes made by the six radiologists from “incorrect” assessments to “correct” ones because of the CAD system’s influence ranged from 0% (0 of 124) to 14.5% (18 of 124). For benign nodules, the correct changes ranged from 0% (0 of 132) to 5.3% (7 of 132). The impact of CAD therefore could be substantial for some radiologists but minimal for others, even among experienced thoracic radiologists.

The radiologists changed their recommended actions an average of 10.8 ± 5.8 times (range, 5–18 times). We considered a change to be “correct” for a malignant nodule when the recommended action was changed from “no action” to either “CT follow-up” or “immediate action,” or “CT follow-up” was changed to “immediate action,” and vice versa for an “incorrect” change. Correct recommended action changes were made an average of 6.8 ± 2.5 times (range, 4–10 times), while an average of 4 ± 3.6 incorrect recommended action changes (range, 1–9 changes) were made. These changes are also summarized in Table 2. The change in the recommended action, however, was not statistically significant for any reader for either malignant or benign nodules by McNemar’s test.

We analyzed observer performance on two subsets of the data (Table 3): (1) primary cancers and benign nodules, and (2) metastatic cancers and benign nodules. For the primary cancer subset, the average A_z of the radiologists improved significantly ($P < .01$) from 0.823 (range, 0.805–0.837) without CAD to 0.848 (range, 0.823–0.866) with CAD. Their average $A_z^{0.9}$ improved significantly from 0.338 to 0.415 ($P = .045$). For the metastatic cancer subset, the average A_z of the radiologists also improved from 0.849 (range, 0.813–0.877) without CAD to 0.861 (range, 0.834–0.895) with CAD, but the improvement fell short of statistical significance ($P = .06$). Their average $A_z^{0.9}$ improved significantly from 0.493 to 0.535 ($P = .01$).

For nodule feature assessment, there was good agreement in some but considerable interobserver variability in other features. If we considered the situation that all radiologists selected the same feature descriptor for a given nodule as having feature agreement, radiologists agreed on the presence or absence of cavitation for 93.4% of the nodules. The agreement for other features, calcification, nodule edge, and attenuation, was not as high, estimated to be 76.6%, 46.5%, and 59.0%, respectively.

There were eight masses with diameters > 30 mm in the data set. If the test scores of the eight masses were removed, the test A_z of the CAD system was 0.849 ± 0.024 . When the scores of the eight masses were removed from each observer’s data, the average A_z for the observers still improved significantly ($P < .01$) from 0.832 (range, 0.813–0.853) without CAD to 0.850 (range, 0.837–0.879) with CAD. The average $A_z^{0.9}$ improved significantly ($P = .024$) from 0.392 (range, 0.311–0.446) without CAD to 0.455 (range, 0.402–0.548) with CAD. These results are also summarized in Table 3.

The effects of CAD on radiologists’ performance are demonstrated by the following examples. Figure 5 is an example of the beneficial influence of CAD. This was a biopsy-proven non-small-cell lung cancer. Radiologists gave an average LM of 45.8% (range, 5%–70%) without CAD but increased it to 58.3% after seeing the classifier score of 7. Figure 6

is another example of the beneficial influence of CAD. This nodule was determined to be benign after no changes were observed over 2 years. Radiologists gave an average LM of 53.3% (range, 20%–65%) without CAD but reduced it to 48.3% after seeing the classifier score of 4. Figure 7 is an example for which the CAD system gave an incorrect score that did not adversely affect radiologists. This nodule was found to be adenoid cystic carcinoma by biopsy. Radiologists gave an average LM of 57.5% (range, 35%–90%), but they did not modify their ratings substantially after seeing the classifier score of 4, as the average LM with CAD was 55.8% (range, 35%–90%).

DISCUSSION

Our results indicate that CAD can benefit radiologists in characterizing lung nodules on CT images. The improvement was modest, though significant, possibly because the radiologists participating in the study are experienced fellowship-trained thoracic radiologists. Experience level could influence how beneficial CAD is for a radiologist. Awai et al (27) reported that radiology residents showed significant improvement, from an average A_z of 0.768 ± 0.078 to 0.901 ± 0.036 ($P = .009$), but there was no significant improvement for board-certified radiologists.

As with any other second opinion, a radiologist may or may not concur with the suggestion and may even change for the worse on second thought. The radiologists were cautious in making changes, as the classifier scores affected them enough to modify their malignancy assessments for only half the nodules on average. Even when the CAD system assessment led them to make changes, they modified their scores by an average of only 10 points. This could be attributed to several reasons. First, the observers were all thoracic radiologists experienced in chest CT interpretation. Second, this study was the first experience of using CAD for all observers, so they might not have had strong confidence in the CAD system. Finally, the opinion of a CAD system may not be as effective that of a second radiologist because it is not interactive. In our implementation, the CAD system only provided a relative malignancy rating between 1 and 10 in reference to the rating distributions for the data set. In a consensus double-reading setting, a radiologist can interact with a second radiologist to understand his or her opinion and reasons for arriving at a diagnosis, but this is not possible with CAD. It may be beneficial for a CAD system to provide examples of similar nodules with known diagnoses to justify its decision, as in a content-based image-retrieval CAD system (40). The effectiveness of these two CAD approaches will warrant comparison in future studies.

Our CAD system extracts a variety of features to analyze the nodules. The diversity of selected features, including morphologic, texture, and gradient field features, indicates that many different characteristics of a nodule contain useful information about its malignancy. In particular, the size, edge, and texture surrounding a nodule have been found to be effective discriminators. This is consistent with the findings by other investigators in characterization of nodules (41). The advantage of computerized image analysis is that it can extract features such as the skewness of gray-level histogram or texture descriptors that may not be readily visualized, thus allowing the computer-extracted information to complement the radiologist's assessment. In clinical practice, radiologists use other patient information, such as age, gender, and smoking history, to make diagnoses. In a different study, we compared the performances of the CAD system without and with patient age and gender as available features. Age was selected consistently as an input feature, but the improvement in the test A_z value was <0.01 for this data set (28). Because the gain in performance is minimal and the chance of erroneous or missing input information may increase if CAD is recommended for use in a screening setting, we did not include patient information in the current CAD system.

We included both primary and metastatic cancers in our data set. Primary cancers were the majority of the malignant nodules (72 of 124 [58%]). These are the cancers that would be of main concern, because CAD would be beneficial when primary cancers are correctly characterized in otherwise asymptomatic patients. There was significant improvement ($P < .05$) in radiologists' classifications of primary cancers with CAD. Metastatic cancers were included in our data set because they would also appear on clinical scans, although radiologists may be more vigilant and suspicious of metastatic diseases in patients with other cancers. CAD also improved the radiologists' characterization accuracy for metastatic cancers, but it fell short of statistical significance.

In our observer study, the entire CT scan was available to the observer, although the observer was instructed to make an assessment on the basis of the characteristics of the nodule of interest. The surrounding lung parenchymal features or the presence or absence of other nodules in the scan would potentially be used by the observer to reach the diagnostic decision. In comparison to an approach in which a cropped VOI containing only the nodule of interest would be shown to an observer, our approach would be more similar to reading in clinical practice and might reduce the chance of an optimistic bias on the observed effect of CAD; without the surrounding parenchymal information, the radiologist might be less confident in his or her decision and more likely to be influenced by the CAD system's rating. The possible correlation between the rating data of nodules from the same scan on the ROC analysis was corrected by using the Obuchowski method for analysis of clustered data (37).

One issue that concerns all CAD researchers is the lack of a large training and independent test set. It is difficult to reserve an independent set for testing because of the countering priority to have as many samples as possible for training. Before larger sets from the Lung Imaging Database Consortium or other sources are available, resampling schemes such as the leave-one-out method may be used to estimate test performance. Our study used such an approach to produce test scores for the nodule set. The results of the ROC study indicate that, given the use of a CAD system with the level of performance as that in this study, the observers could achieve significantly higher accuracy with CAD than without CAD. This relative improvement demonstrates the potential benefit of using CAD.

Large interobserver and intraobserver variability has been reported on such tasks as the estimation of LM or segmentation (42,43). We asked the radiologists to rate nodule features in this study. Considerable interobserver variability was observed among the nodule feature descriptors provided by radiologists. This demonstrates that the radiologists may have perceived differences in whether a nodule had solid or ground-glass opacity components, for example. Because features such as the margin or the presence of calcification are also useful indicators of the LM of a nodule (41), when radiologists do not agree on their perception of these features, it may follow that radiologists' assessments of malignancy would be substantially different. Although the CAD system did not change the radiologists' perception of the features, it still helped reduce slightly the interobserver variability in the LM estimates. The standard deviation of the LM ratings from the six radiologists for the individual nodules decreased significantly ($P = .0024$) from an average of 15.34 without CAD to an average of 14.54 using a paired t test. The reduction in interobserver variability by CAD was discussed by Jiang et al (44).

CAD systems for detecting lung nodules have been approved by the US Food and Drug Administration and are already in clinical use. CAD systems for characterization have not been commercially available, but research is ongoing. Computerized image analysis by a CAD system will likely provide useful information complementary to, but not in place of, the diagnostic and clinical information that radiologists routinely use for the assessment of

nodule malignancy. Radiologists should be well informed of the performance of the specific CAD system before implementing it for clinical practice. They should also evaluate the CAD system over time on the basis of their own experience in assessing nodules without and with CAD. Only if they understand the benefits and limitations can they take best advantage of the information provided by the CAD system and use it properly as a second opinion.

There were limitations in our study. The participating readers were fellowship-trained thoracic radiologists and did not accurately reflect the population of radiologists in general. Thus, these results should not be extrapolated to the performance of radiologists as a whole, though it is encouraging to observe that the A_z performance of all six experienced radiologists in our study improved with CAD. We plan to conduct an observer study for radiologists who vary in experience to evaluate the effect of CAD on their diagnostic performance.

A second limitation is the fact that this study was a controlled laboratory experiment, in which radiologists knew that they would be reading cases containing nodules in succession. The prevalence of lung nodules on CT imaging, especially malignant ones, was not reflective of what radiologists would typically see in clinical practice. Because their assessments of the nodules in this retrospective observer study would not affect patient care, there is a possibility that their response to the second opinion by CAD could be different from that in an actual clinical setting. The laboratory effect was discussed by Gur et al (45). To evaluate the clinical significance of CAD, future prospective studies with independent cases will be needed to determine whether CAD can influence radiologists' patient management decisions in clinical practice and how it will affect patient care. Nevertheless, a laboratory ROC experiment is generally a first step to demonstrate the feasibility of a CAD system before a costly clinical trial is performed.

Third, the small number of changes in recommended action could have been due to the very broad range of options each choice encompassed. For example, 3-month, 6-month, and 12-month follow-up all fell under "CT follow-up," while "immediate action" could have included sputum analysis, positron emission tomography, or surgery. A change of <20% in the LM estimates therefore might not make a difference in the recommended action in most cases. Furthermore, even if a CAD system is highly accurate, radiologists will have to develop strong confidence through their experience with the CAD system before they will be willing to change a recommended action in clinical situations, because of medicolegal issues. Until CAD is deployed in a real-world situation, it is impossible to know whether it would be truly beneficial.

The fourth limitation is that we used a heterogeneous data set with a wide range of scan parameters, including slice thickness, slice interval, and dose. The variations in the scan parameters may increase the variations in the image features, thereby reducing the differentiation of the malignant and benign nodules. We did not use a data set with more homogeneous scanning parameters, because of the limited availability of cases with known diagnoses. Furthermore, if the CAD system and the extracted features are tailored to suit only a data set of homogeneous imaging parameters or thin-slice scans, the estimated performance will likely be overly optimistic compared to what will be achieved if our method and features are applied to data sets acquired with different parameters. Given the large intrainstitution and interinstitution variations in scan parameters and image quality in a clinical environment, the use of a mixed data set as in our study may provide a more realistic estimate of an average performance when the CT scan is not of ideal image quality.

Finally, we have taken steps to prevent overtraining and bias in our classifier by using the two-loop leave-one-case-out training and testing method so that there is an independent test case for each training cycle. However, we are aware that there is a possibility of inadvertent overtraining by virtue of using the same data set many times to adjust the CAD system parameters. We will continue to expand the database from our patient files and also expect to use the Lung Imaging Database Consortium public data set if pathologic results of the lung nodules are made available in the future.

In conclusion, we performed an observer study to evaluate the effects of CAD on the diagnostic performance of fellowship-trained experienced thoracic radiologists for lung nodules on CT imaging. Our data set included both primary and metastatic lung cancers, and we analyzed the assessments of the two groups collectively and separately. We found that radiologists obtained significant improvement in diagnostic accuracy with the computer aid. The recommended action changes as a result of CAD were also mostly beneficial. These results suggest that CAD may be helpful as a second opinion in increasing diagnostic confidence for radiologists. Future work includes expanding the data set to increase the number of training samples, improving the performance of the CAD system, and evaluating the CAD system with a previously unseen test set. Further studies are also needed to determine whether similar improvement in diagnostic accuracy will be realized for radiologists of different experience levels and whether CAD can influence radiologists' patient management decisions and affect patient care in clinical practice.

Acknowledgments

This work was supported by grant CA93517 from the US Public Health Service (Rockville, MD).

We are grateful to Charles E. Metz, PhD, for the LABMRC program.

REFERENCES

1. Sone S, Takashima S, Li F, et al. Mass screening for lung cancer with mobile spiral computed tomography scanner. *Lancet*. 1998; 351:1242–1245. [PubMed: 9643744]
2. Henschke CI, McCauley DI, Yankelevitz DF, et al. Early lung cancer action project: overall design and findings from baseline screening. *Lancet*. 1999; 354:99–105. [PubMed: 10408484]
3. Nawa T, Nakagawa T, Kusano S, et al. Lung cancer screening using low-dose spiral CT: results of baseline and 1-year follow-up studies. *Chest*. 2002; 122:15–20. [PubMed: 12114333]
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. [PubMed: 11867800]
5. Henschke CI, Yankelevitz DF, Libby DM, et al. Survival of patients with stage I lung cancer detected on CT screening. *N Engl J Med*. 2006; 355:1763–1771. [PubMed: 17065637]
6. Sobue T, Moriyama N, Kaneko M, et al. Screening for lung cancer with low-dose helical computed tomography: anti-lung cancer association project. *J Clin Oncol*. 2002; 20:911–920. [PubMed: 11844811]
7. Gierada DS, Pilgram TK, Ford M, et al. Lung cancer: interobserver agreement on interpretation of pulmonary findings at low-dose CT screening. *Radiology*. 2008; 246:265–272. [PubMed: 18024436]
8. Swensen SJ, Jett JR, Hartman TE, et al. Lung cancer screening with CT: Mayo Clinic experience. *Radiology*. 2003; 226:756–761. [PubMed: 12601181]
9. Das M, Muhlenbruch G, Mahnken AH, et al. Small pulmonary nodules: Effect of two computer-aided detection systems on radiologist performance. *Radiology*. 2006; 241:564–571. [PubMed: 17057074]
10. Brown MS, Goldin JG, Rogers S, et al. Computer-aided lung nodule detection in CT results of large-scale observer test. *Acad Radiol*. 2005; 12:681–686. [PubMed: 15935966]

11. Awai K, Murao K, Ozawa A, et al. Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists' detection performance. *Radiology*. 2004; 230:347–352. [PubMed: 14752180]
12. Armato S, Li F, Giger M, et al. Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program. *Radiology*. 2002; 225:685–692. [PubMed: 12461246]
13. Li F, Arimura H, Suzuki K, et al. Computer-aided detection of peripheral lung cancers missed at CT: ROC analyses without and with localization. *Radiology*. 2005; 237:684–690. [PubMed: 16244277]
14. Sahiner B, Hadjiiski LM, Chan HP, et al. The effect of nodule segmentation on the accuracy of computerized lung nodule detection on CT scans: comparison on a data set annotated by multiple radiologists. *Proc SPIE*. 2007; 6514:65140L–65141L. 65147.
15. Way TW, Hadjiiski LM, Sahiner B, et al. Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. *Med Phys*. 2006; 33:2323–2337. [PubMed: 16898434]
16. Shah SK, McNitt-Gray MF, Rogers SR, et al. Computer-aided diagnosis of the solitary pulmonary nodule. *Acad Radiol*. 2005; 12:570–575. [PubMed: 15866129]
17. McNitt-Gray MF, Hart EM, Wyckoff N, Sayre JW, Goldin JG, Aberle DR. A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results. *Med Phys*. 1999; 26:880–888. [PubMed: 10436888]
18. Armato SG, Altman MB, Wilkie J. Automated lung nodule classification following automated nodule detection on CT: a serial approach. *Med Phys*. 2003; 30:1188–1197. [PubMed: 12852543]
19. Kawata Y, Niki N, Ohmatsu H, et al. Quantitative surface characterization of pulmonary nodules based on thin-section CT images. *IEEE Trans Nucl Sci*. 1998; 45:2132–2138.
20. Lo SCB, Hsu LY, Freedman MT, et al. Classification of lung nodules in diagnostic CT: an approach based on 3-D vascular features, nodule density distributions, and shape features. *Proc SPIE*. 2003; 5032:183–189.
21. Li F, Aoyama M, Shiraishi J, et al. Radiologists' performance for differentiating benign from malignant lung nodules on high-resolution CT using computer-estimated likelihood of malignancy. *AJR Am J Roentgenol*. 2004; 183:1209–1215. [PubMed: 15505279]
22. Aoyama M, Li Q, Katsuragawa S, et al. Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images. *Med Phys*. 2003; 30:387–394. [PubMed: 12674239]
23. Suzuki K, Li F, Sone S, Doi K. Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network. *IEEE Trans Med Imaging*. 2005; 24:1138–1150. [PubMed: 16156352]
24. 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 Am J Roentgenol*. 2002; 178:657–663. [PubMed: 11856693]
25. Shah SK, McNitt-Gray MF, De Zoysa KR, et al. Solitary pulmonary nodule diagnosis on CT results of an observer study. *Acad Radiol*. 2005; 12:496–501. [PubMed: 15831424]
26. Li F, Li Q, Engelmann R, et al. Improving radiologists' recommendations with computer-aided diagnosis for management of small nodules detected by CT. *Acad Radiol*. 2006; 13:943–950. [PubMed: 16843846]
27. Awai K, Murao K, Ozawa A, et al. Pulmonary nodules: Estimation of malignancy at thin-section helical CT—effect of computer-aided diagnosis on performance of radiologists. *Radiology*. 2006; 239:276–284. [PubMed: 16467210]
28. Way TW, Sahiner B, Chan H-P, et al. Computer aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. *Med Phys*. 2009; 36:3086–3098. [PubMed: 19673208]
29. Sahiner B, Chan HP, Petrick N, et al. Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. *Med Phys*. 1998; 25:516–526. [PubMed: 9571620]

30. Galloway MM. Texture classification using gray level run lengths. *Comput Graph Image Process.* 1975; 4:172–179.
31. Dasarathy BR, Holder EB. Image characterizations based on joint gray-level run-length distributions. *Patt Recog Lett.* 1991; 12:497–502.
32. Ge Z, Sahiner B, Chan HP, et al. Computer aided detection of lung nodules: false positive reduction using a 3D gradient field method and 3D ellipsoid fitting. *Med Phys.* 2005; 32:2443–2454. [PubMed: 16193773]
33. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol.* 1989; 24:234–245. [PubMed: 2753640]
34. 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. [PubMed: 1399456]
35. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology.* 1996; 201:745–750. [PubMed: 8939225]
36. Sahiner B, Chan HP, Petrick N, et al. Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis. *Phys Med Biol.* 1998; 43:2853–2871. [PubMed: 9814523]
37. Obuchowski N. Nonparametric analysis of clustered ROC curve data. *Biometrics.* 1997; 53:567–578. [PubMed: 9192452]
38. Lee M-LT, Rosner BA. The average area under correlated receiver operating characteristic curves: a nonparametric approach based on generalized two-sample Wilcoxon statistics. *J R Stat Soc Ser C Appl Stat.* 2001; 50:337–344.
39. Hadjiiski LM, Chan HP, Sahiner B, et al. Breast masses: computer-aided diagnosis with serial mammograms. *Radiology.* 2006; 240:343–356. [PubMed: 16801362]
40. Li Q, Li F, Shiraishi J, Katsuragawa S, et al. Investigation of new psychophysical measures for evaluation of similar images on thoracic computed tomography for distinction between benign and malignant nodules. *Med Phys.* 2003; 30:2584–2593. [PubMed: 14596294]
41. Gurney JW. Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis—part I. Theory. *Radiology.* 1993; 186:405–413.
42. Bogot NR, Kazerooni EA, Kelly AM, et al. Interobserver and intraobserver variability in the assessment of pulmonary nodule size on CT using film and computer display methods. *Acad Radiol.* 2005; 12:948–956. [PubMed: 16087090]
43. Meyer CR, Johnson TD, McLennan G, et al. Evaluation of lung MDCT nodule annotation across radiologists and methods. *Acad Radiol.* 2006; 13:1254–1265. [PubMed: 16979075]
44. Jiang Y, Nishikawa RM, Schmidt RA, et al. Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. *Radiology.* 2001; 220:787–794. [PubMed: 11526283]
45. Gur D, Bandos AI, Cohen CS, et al. The “laboratory” effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology.* 2008; 249:47–53. [PubMed: 18682584]

Figure 1.
Histograms of the longest diameters of the benign and malignant nodules as measured by experienced chest radiologists.

Figure 2.
The 10-bin histogram of classifier scores with fitted Gaussian distributions for the malignant and benign classes.

a.

b.

Figure 3.

(a) The graphical user interface used by radiologists in the observer study. The first slice of a scan presented is the one containing the nodule marked in a box. (b) The CAD system score that would appear in the upper middle of the screen after the user clicks “Load CAD.”

Figure 4.

The averaged ROC curves for the six radiologists without ($A_z=0.834$) and with CAD ($A_z=0.854$) ($p<0.01$) derived from the average slope and intercept parameters of the individual observers' fitted ROC curves, and the CAD system performance (test $A_z=0.857\pm0.023$). CAD, computer-aided diagnosis.

Figure 5.
Example of a non-small cell lung cancer that radiologists gave an average likelihood of malignancy of 45.8%, but increased it to 58.3% after seeing the classifier rating of 7, showing the beneficial effect of CAD.

Figure 6. Example of a benign nodule that radiologists gave an average likelihood of malignancy of 53.3%, but was reduced to 48.3% after seeing the classifier score of 4, showing the beneficial effect of CAD.

Figure 7.

Biopsy determined that this was adenoid cystic carcinoma, and radiologists gave it an average LM of 57.5%. Though the classifier score of 4 was incorrect, the radiologists changed the likelihood to an average of 55.8%, showing that radiologists are not easily misled by the CAD system if they believe CAD is incorrect.

TABLE 1Individual and Average Performance of the Observers in Terms of A_z without and with CAD

Observer	A_z without CAD	A_z with CAD
1	0.817 ± 0.026	0.846 ± 0.024
2	0.845 ± 0.025	0.857 ± 0.024
3	0.843 ± 0.024	0.847 ± 0.024
4	0.829 ± 0.025	0.853 ± 0.023
5	0.847 ± 0.024	0.877 ± 0.021
6	0.817 ± 0.026	0.834 ± 0.025
Average *	0.833	0.854

A_z , area under the receiver-operating characteristic curve; CAD, computer-aided diagnosis.

* The average A_z was obtained as the mean of the six individual A_z values. The improvement in the average A_z achieved statistical significance ($P < .01$).

TABLE 2

The Number of Times on Average That Observers Changed Their LM Estimates and Recommended Actions with CAD

	LM Change	Action Change
Total	126 ± 46.8 (49 ± 17%)	10.8 ± 5.8
Correct	95 ± 34.0 (37 ± 13%)	6.8 ± 2.5 (3 ± 1%)
Incorrect	31 ± 18.2 (12 ± 7%)	4 ± 3.6 (2 ± 1%)

CAD, computer-aided diagnosis; LM, likelihood of malignancy.

Values in parentheses are percentages relative to the total number of nodules.

TABLE 3

Average Performance of the Observers in Terms of A_z and $A_z^{0.9}$ for the Entire Data Set, the Primary and Metastatic Subsets, and the Data Set Excluding the Eight Masses > 30mm in Diameter

Data Set	Average A_z		Average $A_z^{0.9}$	
	Without CAD	With CAD	Without CAD	With CAD
All 256 nodules	0.833	0.853	0.390	0.456
Primary cancers	0.823	0.846	0.338	0.415
Metastatic cancers	0.849	0.861	0.493	0.535
Excluding masses > 30 mm in diameter	0.832	0.850	0.392	0.455

A_z , area under the receiver-operating characteristic curve; $A_z^{0.9}$, partial A_z (the area under the receiver-operating characteristic curve above a true-positive fraction value of 0.9) CAD, computer-aided diagnosis.

All improvements with CAD were statistically significant ($P < .05$), except for the metastatic cancer subset ($P = .06$).