Improving Radiologists' Recommendations With Computer-Aided Diagnosis for Management of Small Nodules Detected by CT[1]

Feng Li, Qiang Li, Roger Engelmann, Masahito Aoyama, Shusuke Sone, Heber MacMahon, Kunio Doi

Rationale and Objectives. To evaluate how computer-aided diagnosis (CAD) can improve radiologists' recommendations for management of possible early lung cancers on CT.

Materials and Methods. Twenty-eight lung cancers and 28 benign lesions were employed. Each group of 28 lesions was classified into subgroups of two sizes (9 between 6 and 10 mm and 19 between 11 and 20 mm) and three patterns (8 with pure ground glass opacity [GGO], 12 with mixed GGO and 8 solid lesions). Sixteen radiologists participated in the observer study, first without and then ,with CAD. Radiologists' recommendations, including (1) follow-up in 12 months, (2) in 6 months, (3) in 3 months, or (4) biopsy, were compared at three levels of their malignancy probability ratings (low: 1%–33%; medium: 34%–66%; high: 67%–99%) for 896 observations (56 lesions by the 16 radiologists) in the two size subgroups and three patterns.

Results. The number of recommendations changed by radiologists by use of CAD was 163 (18%) among all 896 observations. Among these changed recommendations, the fraction showing a beneficial effect from CAD was 68% (111/163), and the fraction showing a beneficial effect regarding biopsy recommendations was 69% (48/70). With CAD, the radiologists' performance regarding biopsy recommendations was significantly improved for 43 lung cancers (31 changed to biopsy versus 12 changed away from biopsy; $P = .003$) and was also improved for 27 benign lesions (10 changed to biopsy versus 17 changed away from biopsy; $P = .18$). Most of the cancers with improved recommendations were solid lesions or mixed GGO and relatively large.

Conclusion. CAD has the potential to improve the appropriateness of radiologists' recommendations for small malignant and benign lesions on CT scans.

Key Words. Lung neoplasms, CT; Computer diagnostic aid; Lung, module.

Acad Radiol 2006; 13:943–950

[1] From the Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, The University of Chicago, Chicago, IL (F.L., Q.L., R.E., H.M., K.D.); Department of Intelligent Systems, Faculty of Information Sciences, Hiroshima City University, Hiroshima, Japan (M.A.); and JA Nagano Chu-Nanshin Regional Cancer Screening Centre, Ikeda, Nagano, Japan (S.S.). Received December 13, 2005; accepted April 17, 2006. This work was partly supported 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. CAD technologies developed in the Kurt Rossmann Laboratories have been licensed to companies including R2 Technology, Deus Technologies, Riverain Medical Group, Mitsubishi Apace Software Co., Median Technologies, General Electric Corporation, and Toshiba Corporation. It is the policy of The University of Chicago that investigators disclose publicly actual or potential significant financial interests that may appear to be affected by research activities.**Address correspondence to:** F.L. e-mail: feng@uchicago.edu

© AUR, 2006
doi:10.1016/j.acra.2006.04.010

Among diagnostic imaging modalities, computed tomography (CT) has the highest sensitivity for detection of small pulmonary lesions. However, it is difficult for radiologists to correctly distinguish cancers from noncancerous lesions (false positives) and to make appropriate and consistent recommendations management of patients with suspicious lesions. On the one hand, a large number of false positives will lead to unnecessary patient anxiety and will increase the increased economic costs and radiation exposure. A high rate of false positives can also lead to unnecessary investigation such as CT scans, biopsy, and even surgery. On the other hand, in the case of lung cancers (true positives), if radiologists fail to make an appropriate recommendation such as biopsy or surgery, the patients may miss an opportunity for cure.

The Food and Drug Administration has approved the clinical use of some computer-aided diagnosis (CAD) detection systems in screening for clinical use, especially for breast cancer screening on mammography in the United States. Gur et al (1) reported that the introduction of detection CAD into a large clinical practice (115,571 screening mammograms) was not associated with statically significant changes in both recall and breast cancer detection rates. Commercially available detection CAD systems show marks, including true positives (cancers) and false positives (noncancerous lesions also anatomic structures), on each whole image (1–3). Recently, automatic classification CAD schemes for distinction of malignant and benign lesions have been developed in some universities (4–8) that show an estimated likelihood of malignancy for each segmented lesion based on its image features. Some observer studies using mammograms reported that classification CAD had a beneficial effect for radiologists' diagnostic accuracy for classifying malignant and benign breast masses and their recommendations regarding biopsy (5,6).

It is important that a larger database, including large number of lesions and a variety of lesion patterns, be used for developing classification CAD. The thin-section CT database for developing our classification CAD scheme used in this study comprised follow-up exams obtained from a 3-year CT lung cancer screening program (17,892 examinations). The database included 61 primary lung cancers (size range 6–19 mm; mean 12 mm) and 183 benign nodules (size range 3–20 mm; mean 7 mm) with three different patterns (8,9). We have reported (8) that our CAD scheme has the potential to improve radiologists' diagnostic accuracy for lesion classification and also to improve radiologists' recommendations in an observer study. The data analysis in the previous report (8) was independently calculated for 16 observers, and the radiologists' recommendations were improved by increasing the number of biopsy recommendations for actual early cancers (statistically significant) and by reducing the number for actual benign ones (not significant) in an observer study. The current study used the same data from the same observer test as used previously (8). Our purpose in this study was to evaluate further how CAD can assist radiologists in their recommendation management of possible early lung cancers that have different sizes and patterns.

MATERIALS AND METHODS

Institutional review board approval and informed observer consent were obtained.

Database

Our database was obtained as part of an annual 3-year CT screening for lung cancer in a general population in Nagano, Japan (8,9), which included 59 patients (27 men, 32 women, mean age 64.6 years) with 61 primary small lung cancers (mean size 12.3 mm; size range 6–20 mm; 18 nodules with pure ground glass opacity [GGO]; 28 with mixed GGO; and 15 with solid opacity), and 169 patients (99 men, 70 women, mean age 61.6 years) with 183 benign lesions (mean size 7.2 mm; size range 3–20 mm; 12 with pure GGO, 30 with mixed GGO, and 141 with solid opacity). All patients gave informed consent. All cancers were confirmed by surgery, and benign lesions were confirmed by surgery or follow-up (resolved or no change for 2 years or more). The mean size (average length and width) of each nodule was recorded by one radiologist (F.L.). The three types of patterns of these lesions, including pure GGO, mixed GGO, and solid opacity, were viewed independently and grouped by three radiologists (F.L. among them) without knowledge of the final diagnosis, and then a consensus was reached through discussion. Thin-section CT scans were performed on a helical scanner (CT HiSpeed Advantage, GE, Milwaukee, WI) with a standard tube current (200 mA) to cover the entire lesion, with 1-mm collimation and a bone reconstruction algorithm with a 0.5-mm interval.

CAD

With our CAD scheme, the nodules were segmented automatically by use a dynamic programming technique.

The technique has been described in detail elsewhere (7). A total of 41 and 15 image features based on two-dimensional and three-dimensional volume data, respectively, were determined from quantitative analysis of the nodule outline and pixel values. Linear discriminant analysis was employed for distinguishing benign from malignant nodules. The performance of this CAD scheme was evaluated based on a "leave-one-out" testing method by use of 61 malignant and 183 benign nodules. For the input of the linear discriminant analysis, we selected many combinations from 56 features and two clinical parameters (age and gender). The final features included effective diameter, contrast, margin or edge, shape, attenuation, and internal homogeneity of the segmented nodules.

Our computerized classification method outputs a percentage (1%–99%) indicating the likelihood of malignancy. The performance of the classification scheme yielded an A_z value of 0.937 (0.934 for lesions at 6–10 mm, 0.855 for lesions at 11–20 mm, 0.919 for nodules with pure GGO, 0.852 for nodules with mixed GGO, and 0.957 for solid nodules) for distinction between 61 lung cancers and 183 benign nodules.

Case Selection

Twenty-eight patients (mean age 63.4 years; 14 men and 14 women) with lung cancers and 28 patients (mean age 64.2 years; 17 men and 11 women) with benign lesions on thin-section CT were included in this observer study. The 28 malignant lesions were randomly selected from 61 lung cancers, and the 28 benign lesions were selected by matching of their size and pattern to the cancers from 183 benign lesions among our database. For both cancers and benign lesions, 9 lesions were in the range of 6–10 mm and 19 lesions in the range of 11–20 mm; the lesion patterns were 8 pure GGO, 12 mixed GGO, and 8 solid opacity. The performance of the classification scheme yielded an A_z value of 0.831 (0.842 for lesions at 6–10 mm, 0.870 for lesions at 11–20 mm, 0.910 for nodules with pure GGO, 0.814 for nodules with mixed GGO, and 0.783 for solid nodules) for the 28 lung cancers and 28 benign nodules. The 56 lesions used in this observer study were the largest number of lesions that could be matched in size and pattern between the 183 benign lesions and the 61 lung cancers in our database.

The 28 cancers included 19 well-differentiated adenocarcinomas, 5 other adenocarcinomas, 2 squamous cell carcinomas, and 2 localized small-cell carcinomas. Among the 28 benign lesions, 2 (inflammatory pseudotumor and sclerosing hemangioma) were confirmed by surgery, 19 had resolved on follow-up examination, and 7 had not changed for 2 years or more.

Observer Study

Sixteen radiologists (H.M. among them) participated in this observer study. The 16 radiologists, including 7 chest radiologists and 9 general radiologists, have a mean of 14 years of experience (range 7–26 years). Consecutive region of interest images for each lesion on thin-section CT were presented for interpretation by use of a cine-type display on a high-resolution CRT monitor. The windowing was initially set at a width of 1500 Hounsfield units and a level of −550 Hounsfield units, but could be adjusted by the observer. In addition, zooming capability was provided. Two clinical parameters (age and gender) were provided to the observers on the monitor.

It was explained to the observers that the purpose of this study was to assist radiologists in distinguishing benign from malignant lesions on thin-section CT by use of a CAD scheme. The observers were informed that the lesions used in this study were obtained from an annual 3-year CT screening for lung cancer in a general population in Japan. The instructions for the observers included (a) the role of CAD output as a "second opinion;" (b) 28 malignant (6–10 mm: 9 cases; 11–20 mm: 19 cases; and pure GGO: 8 cases, mixed GGO: 12 cases, and solid opacity: 8 cases) and 28 benign lesions (matched to the cancers in size and pattern) are included in this study; (c) the sensitivity and specificity of our CAD scheme, for a threshold of 50% likelihood of malignancy, are 80% and 75%, respectively; (d) click on a bar (left: benignancy, right: malignancy) on the screen by using a mouse to indicate your confidence level regarding the likelihood of malignancy (from 1% to 99%) of a lesion first without and then with computer output; and (e) after indicating your confidence (without and with CAD), click on one of four recommendations: (1) return to annual screening, in 12 months; (2) follow-up in 6 months; (3) follow-up in 3 months; or (4) biopsy/surgery.

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. There was no pretest training regarding interpretative guidelines for recommendations to radiologists. Radiologists' recommendations without and with CAD were freely decided by each of the observers in this observer study. The reading time was not limited. The average reading time for 56 test cases by 16 radiologists was 46 minutes (range 28–100 minutes; 0.82 minute per case).

Data Analysis

The radiologists' recommendations without and with CAD were analyzed for 896 observations (56 lesions by the 16 radiologists) and were compared at three levels of malignancy (low: 1%–33%; medium: 34%–66%; and high: 67%–99%) for malignant and benign lesions. The test for proportion was used for comparison of the difference in changes on recommendations between those having a beneficial and those have a detrimental effect from CAD for malignant and benign lesions. A chi-square test for independence was used for comparison of the difference in the proportions between radiologists' biopsy recommendations without and with CAD. The recommendations were further classified as "biopsy" and "other" for highly suspicious lesions for which the radiologists indicated their confidence ratings to be 67%–99%. The chi-square test (including a multiple-group test) was used independently for comparison of the difference between (1) lesion sizes (lesions at 6–10 mm and those at 11–20 mm) and (2) lesion patterns (pure GGO, mixed GGO, and solid opacity) for biopsy recommendations on these highly suspicious lesions, without and with CAD.

RESULTS

Figure 1 shows the correlation between computer output and change in the 16 radiologists' recommendations for 896 observations. With CAD, the fraction by which the radiologists changed their recommendations was 18% (163/896), including 18% (80/448) for cancers and 19% (83/448) for benign lesions. Among these changed recommendations, the fraction having a beneficial effect (malignant: step up; benign: step down) was 68% (111/163), and the fraction having a detrimental effect (malignant: step down; benign: step up) was 32% (52/163) because of CAD (test for proportion, $P < .001$). The fractions having a beneficial effect from CAD were 78% (62/80) and 59 % (49/83) for cancers and benign lesions, respectively.

Among the 62 observations for cancers with a beneficial effect, 31 (50%) were changed from follow-up to a biopsy recommendation by 11 radiologists. Among the 49 observations for benign lesions with a beneficial effect, 17 (35%) were changed from biopsy recommendation to follow-up by 9 radiologists. Figure 2a shows a cancer in which the CAD helped four radiologists to improve their recommendation from follow-up to biopsy. Figure 2b shows a benign lesion in which the CAD helped four radiologists to improve their recommendation from biopsy to follow-up.

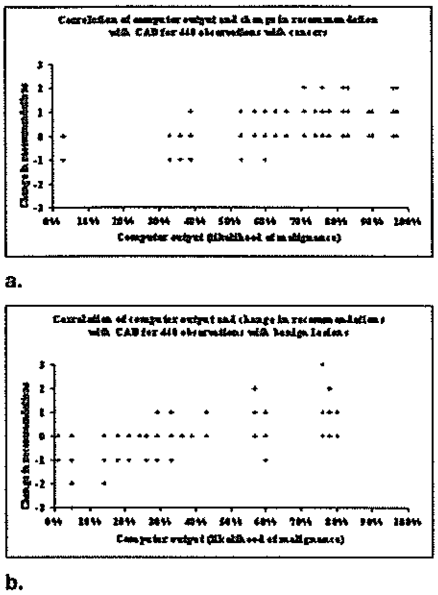


Figure 1. Graphs show the correlation between computer output and change in recommendations for 448 observations (28 cancers by 16 radiologists) **(a)** and 448 observations (28 benign lesions by 16 radiologists) **(b)**. The four recommendation steps are (1) follow-up in 12 months, (2) follow-up in 6 months, (3) follow-up in 3 months, and (4) biopsy. The numbers on the Y axis show the differences in recommendation indices between the without computer-aided diagnosis (CAD) and with CAD conditions: no change (0), step up (1, 2, and 3), and step down (–1, –2, and –3). The number of recommendations changed by radiologists by use of CAD was 163 (18%) for all 896 observations. Among these changed recommendations, the fraction toward a beneficial effect (malignant: step up; benign: step down) because of CAD was 68% (111/163) ($P < .001$).

Table 1 lists the number of lesions grouped based on radiologists' confidence ratings at three levels and recommendations in four steps for 896 observations (448 malignant and 448 benign) without and with CAD. There was no statistical significance in the biopsy recommendations between radiologists without and with CAD for cancers (38% = 170/448 versus 42% = 189/448; P = .22), although the number was increased from 170 to 189. For benign lesions, there was also no statistical significance in biopsy recommendations (13% = 57/448 versus 11% = 50/448; P = .54). The results indicate that the effect was not significant in the total proportion of radiologists' recommendations regarding biopsy by use of CAD.

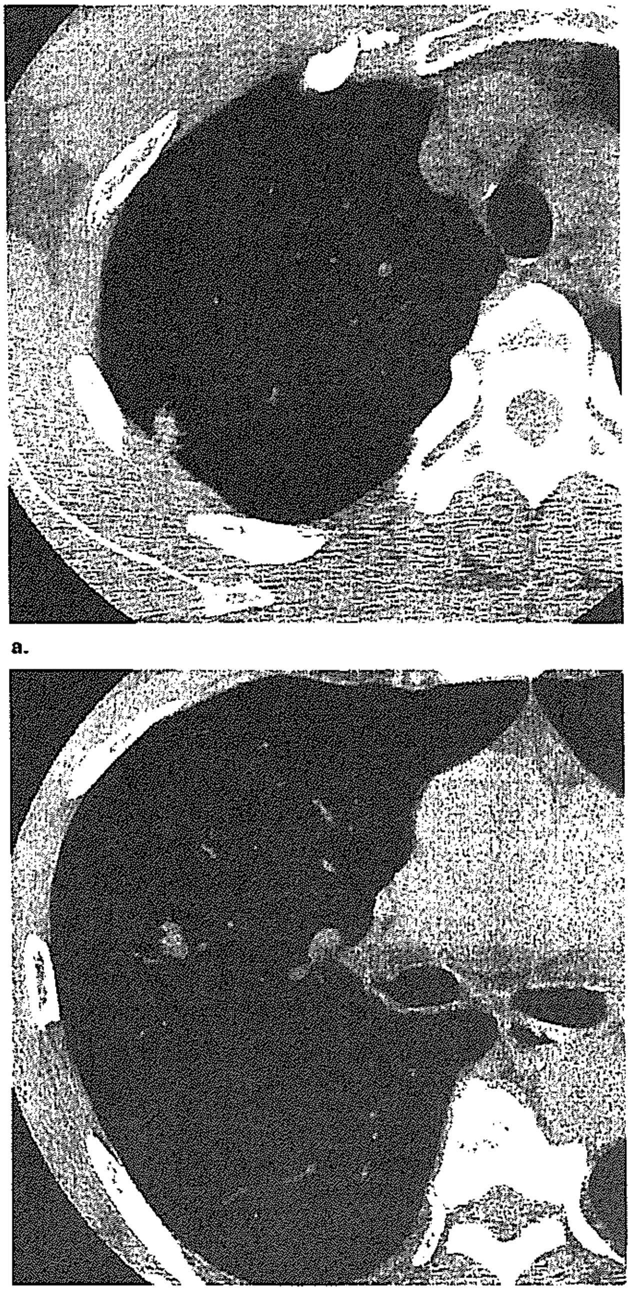


Figure 2. Thin-section computed tomography images in two patients. (a) Computer-aided diagnosis (CAD) (likelihood of malignancy: 71%) helped four radiologists to alter their recommendation from follow-up to biopsy for a 47-year-old man with a squamous cell carcinoma. (b) CAD (likelihood of malignancy: 5%) helped four radiologists to alter their recommendation from biopsy to follow-up for a 63-year-old man with a benign lesion (no change for more than 3 years).

Table 2 shows the distribution of size and pattern of lesions for which radiologists made biopsy recommendations without and with CAD. The difference was statistically significant between a beneficial effect (benign: removed from biopsy; malignant: added to biopsy) and a detrimental effect (benign: added to biopsy; malignant: removed from biopsy) because of CAD (69% = 48/70 versus 31% = 22/70; test for proportion, P = .002). The difference was statistically significant between a beneficial effect and a detrimental effect with CAD for cancers (72% = 31/43 versus 28% = 12/43; P = .003), but the difference was not statistically significant between them with CAD for benign lesions (63% = 17/27 versus 37% = 10/27; P = .18). The results indicate that the changes regarding biopsy recommendations from CAD occurred less frequently for small lesions and lesions with pure GGO.

Table 3 shows the proportion of high confidence ratings (67%–99%) and recommendations for all lesions (malignant and benign) in three subgroups. The difference was statistically significant in the fraction of biopsy recommendations without CAD between lesions at 6–10 mm and lesions at 11–20 mm (32% = 10/31 versus 77% = 158/204; $P < .001$). The difference also was statistically significant for the fraction of biopsy recommendations with CAD between the 6- to 10-mm lesions (31% = 11/35) and the 11- 20-mm lesions (73% = 185/252) ($P <$ .001). The difference was statistically significant in the fraction of biopsy recommendations without CAD within three patterns (pure GGOs: 27% = 12/44; mixed GGOs: 81% = 81/100; and solid lesions: 82% = 75/91; multiple-group test $P < .001$). Further, the difference was statistically significant for the fraction of biopsy recommendations without CAD between pure GGOs and mixed GGOs ($P < .001$) or solid lesions ($P < .001$). There was no statistically significant difference between the mixed GGOs and solid lesions without CAD (P = .95). The difference also was statistically significant for the fraction of biopsy recommendations with CAD within pure GGOs (26% = 16/62), mixed GGOs (78% = 92/118), and solid lesions (82% = 88/107) ($P < .001$), and between pure GGOs and mixed GGOs ($P < .001$) or solid lesions ($P <$ .001). There was no statistically significant difference between the mixed GGOs and solid lesions with CAD (P = .53). The results indicate that radiologists also did not often recommend biopsy for the lesions between 6 and 10 mm and pure GGO lesions even when they indicated a high level of suspicion for cancer, regardless of CAD.

Table 1
Number of Lesions Grouped Based on Three Levels of Radiologists' Confidence and Four Different Recommendations for Lesions Without and With Computer-Aided Diagnosis (CAD)

	Confidence Levels Without CAD				Confidence Levels With CAD			
	1%–33%	34%–66%	67%–99%		1%–33%	34%–66%	67%–99%	
Recommendations	Malignant/Benign			Total	Malignant/Benign			Total
Biopsy	1/0	32/26	137/31	170/57	1/0	22/20	166/30	189/50
Other	79/269	141/113	58/9	278/391	63/287	118/98	78/13	259/398
Follow-up in 12 months	15/98	5/7	0/0	20/105	7/94	3/5	0/0	10/99
Follow-up in 6 months	39/92	48/27	7/1	94/120	29/110	41/28	9/1	79/139
Follow-up in 3 months	25/79	88/79	51/8	164/166	27/83	74/65	69/12	170/160
Total	80/269	173/139	195/40	448/448	64/287	140/118	244/43	448/448

Data are total 896 observations (56 lesions by 16 radiologists), including 448 observations with cancers (28 lesions by 16 radiologists) and 488 observations with benign nodules (28 lesions by 16 radiologists). There was no statistical significance in the biopsy recommendations between radiologists without and with CAD for both cancers (38% = 170/448 versus 42% = 189/448; $P = .22$) and benign lesions (13% = 57/448 versus 11% = 50/448; $P = .54$).

Table 2
Distribution of Lesion Sizes and Patterns for which Radiologists Made Biopsy Recommendations for Lesions Without and With CAD

	Biopsy Recommendations Without CAD	Biopsy Recommendations With CAD	Number (n = 70) With Beneficial (Detrimental) Effect From CAD
	Malignant/Benign	Malignant/Benign	Malignant (n = 43)/Benign (n = 27)
Total	170/57	189/50	31 (12)/17 (10)
Size			
6- to 10-mm lesion	12/5	13/1	3 (2)/4 (0)
11- to 20-mm lesion	158/52	176/49	28 (10)/13 (10)
Pattern			
Pure GGO	14/3	15/4	4 (3)/0 (1)
Mixed GGO	82/20	89/11	13 (6)/10 (1)
Solid	74/34	85/35	14 (3)/7 (8)

CAD, computer-aided diagnosis; GGO, ground glass opacity.

The difference was statistically significant between beneficial effect (malignant: 31 added to biopsy; benign: 17 removed from biopsy) and detrimental effect (malignant: 12 removed from biopsy; benign: 10 added to biopsy) with CAD (69% = 31 + 17/70 versus 31% = 12+10/70; $P < .002$). The difference was statistically significant between a beneficial effect and a detrimental effect with CAD for cancers (72% = 31/43 versus 28% = 12/43; $P = .003$), but the difference was not statistically significant between them with CAD for benign lesions (63% = 17/27 versus 37% = 10/27; $P = .18$). Also the results indicate that the changes regarding biopsy recommendations due to CAD were less occurred for small lesions and lesions with pure GGO.

DISCUSSION

Radiologists' recommendations with use of CAD have been investigated in several observer studies (3,5,6). Some studies showed that there was a significant beneficial effect resulting from classification CAD by increasing biopsy recommendations for breast cancers (5,6) with reduction (6) or no significant change in biopsy recommendations (5) for benign masses. In these studies, no further details were given for the effect of CAD on radiologists' recommendations—for example, how CAD affected radiologists' recommendations concerning different lesion sizes or patterns and why radiologists changed their recommendations for some lesions, but not others.

Observer studies with pulmonary nodules indicated similar results for the improvement of radiologists' performance in detecting lesions and distinguishing benign from malignant lesions on chest radiographs (10–12) and on chest CT scans (8,13–16). In our recent CT studies, we also asked radiologists to indicate their recommendations after they detected suspicious lung lesions (16) or after they had classified small lesions as

Table 3
Proportion of High Confidence Ratings (67%–99%) and Recommendations for All Lesions (Malignant and Benign) in Three Subgroups

	Biopsy Recommendations Without CAD	Biopsy Recommendations With CAD
Size		
6- to 10-mm lesion	32% (10/31)	31% (11/35)
11- to 20-mm lesion	77% (158/204)	73% (185/252)
Pattern		
Pure GGO	27% (12/44)	26% (16/62)
Mixed GGO	81% (81/100)	78% (92/118)
Solid lesion	82% (75/91)	82% (88/107)

CAD, computer-aided diagnosis; GGO, ground glass opacity.
The difference was statistically significant regarding the fraction of biopsy recommendations between the 6- to 10-mm lesions and the 11- to 20-mm lesions without (32% versus 77%; $P < .001$) and with (31% versus 73%; $P < .001$) CAD, and between pure GGOs and mixed GGOs or solid lesions without (27% versus 81% or 82%; $P < .001$) and with (26% versus 78% or 82%; $P < .001$) CAD. The results indicate that radiologists did not often recommend biopsy for small lesions and lesions with pure GGO, even when their level of suspicion for cancer was high, regardless of CAD.

malignant or benign (8). The results indicated that our detection CAD scheme significantly improved radiologists' recommendations for small-cell lung cancers without any significant detrimental effect for false positives on thick-section CT (16). Our classification CAD scheme also significantly improved radiologists' recommendations for early lung cancers, without any significant detrimental effect for small benign lesions on thin-section CT (8). Our purpose in this study was to further evaluate how classification CAD can assist radiologists in improving their recommendations for two sizes (6–10 mm and 11–20 mm) and three patterns (pure GGO, mixed GGO, and solid lesion) of early lung cancers compared with benign lesions.

The findings in the previous work indicated that the improvements in radiologists' confidence ratings resulting from CAD were relatively uniform; the average A_z value was improved from 0.785 to 0.853 for all lesions, including from 0.812 to 0.892 for nodules with pure GGO; from 0.819 to 0.863 for nodules with mixed GGO; and from 0.784 to 0.844 for solid nodules (8). However, the results of the current study indicated that the improvement of radiologists' biopsy recommendations resulting from CAD occurred mostly for larger lesions (11–20 mm) and lesions with mixed GGO or solid opacity. In other words, the current study indicated that the changes in biopsy recommendations were often dependent on lesion sizes or patterns. Radiologists' recommendations regarding biopsy were not often changed for smaller lesions or lesions with pure GGO resulting from CAD although the performance of CAD was also good for classification of these lesions.

We did not give any pretest training regarding interpretative guidelines for recommendations to radiologists in this observer study. However, several CT studies regarding the frequency of malignancy in different sizes and patterns, and regarding the growth rates of the cancers in different patterns, have been published previously (17–26). In the past decade, CT has been applied widely for early lung cancer screening (17–25), and radiologists have learned how lesion size and pattern relate to the probability of malignancy, and how histology affects tumor morphology. For example, the frequency of malignancy was very low for lesion sizes smaller than 10 mm in diameter in a screening program (23), and also in a clinical study (26). GGO lesions are more likely to be malignant than are solid ones in CT screening programs for lung cancer (9,24). In Hasegawa's series, almost all of the GGO lesions were slowly growing lung adenocarcinomas and the mean volume-doubling time of tumors with pure GGO was very long (more than 800 days) (25). Recently, guidelines for management of small pulmonary nodules detected on CT scans have been published (27). In the statement from the Fleischner Society (27), biopsy recommendations are only suggested as an option for lesions larger than 8 mm, whereas long follow-up intervals are appropriate for pure GGOs or very small opacities. These data help explain why radiologists in our study did not often recommend biopsy, even when their level of confidence for cancer was high, regardless of CAD, for the smaller and nonsolid lesions. We believe that radiologists' propensity to recommend biopsy may depend on their perception as to whether the lesion, if cancerous, is likely to grow quickly.

The limitations in this study include the small numbers of malignant and benign lesions. However, the dataset was obtained from a lung cancer CT screening program, which included three different CT patterns for both malignant and benign lesions. We believe that it is more difficult for to distinguish small benign lesions from early lung cancers in similar patterns, especially when distinguishing those lesions with GGO. Therefore, we used a special case subset, which included the most difficult cases in differentiating benign from malignant lesions, in this observer study. There was no case bias for malignant lesions because the 28 lung cancers used in this observer study were selected randomly among our database, and only the 28 benign lesions were selected by matching their patterns and sizes to the cancers. Importantly, with our CAD scheme, the radiologists' performance was improved regarding biopsy recommendations for solid lesions or lesions with mixed GGO at relatively larger sizes. CAD has the potential to be useful for improving management of patients with small lung lesions on CT in clinical practice or in lung cancer screening programs.

ACKNOWLEDGMENT

The authors are grateful to Hiroyuki Abe, MD, Hajime Abiru, MD, Kazuto Ashizawa, MD, Ulrich Bick, MD, Akiko Egawa, MD, Takeshi Kobayashi, MD, Peter MacEneaney, MD, Naohiro Matsuyama, MD, Edward Michales, MD, Kazunori Minami, MD, Tomoaki Okimoto, MD, Gregory Scott Stacy, MD, Shuji Sakai, MD, Christopher Straus, MD, Tetsuji Yamaguchi, MD, and Chaotong Zhang, MD, for participating as observers; and to Mrs. Elisabeth Lanzl for improving the manuscript.

REFERENCES

1. Gur D, Jules H, Sumkin JH, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. J Natl Cancer Inst 2004; 96:185–190.
2. Gur D, Wallace LP, Klym AH, et al. Trends in recall, biopsy, and positive biopsy rates for screening mammography in an academic practice. Radiology 2005; 235:396–401.
3. Marx C, Malich A, Facius M, et al. Are unnecessary follow-up procedures induced by computer-aided diagnosis (CAD) in mammography? Comparison of mammographic diagnosis with and without use of CAD. Eur Radiol 2004; 51:66–72.
4. Doi K, Giger ML, Nishikawa RM, et al. Computer-aided diagnosis of breast cancer on mammograms. Breast Cancer 1997; 25:228–233.
5. Huo Z, Giger ML, Vyborny CJ, et al. Breast cancer: effectiveness of computer-aided diagnosis—observer study with independent database of mammograms. Radiology 2002; 224:560–568.
6. Hadjiiski L, Chan HP, Sahiner B, et al. Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an ROC study. Radiology 2004; 233:255–265.
7. 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.
8. 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.
9. Li F, Sone S, Abe H, et al. Malignant versus benign nodules at CT screening for lung cancer: comparison of thin-section CT findings. Radiology 2004; 233:793–798.
10. MacMahon H, Engelmann R, Behlen FM, et al. Computer-aided diagnosis of pulmonary nodules: results of a large-scale observer test. Radiology 1999; 213:723–726.
11. 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.
12. Shiraishi J, Abe H, Engelmann R, et al. Computer-aided diagnosis to distinguish benign from malignant solitary pulmonary nodules on radiographs: ROC analysis of radiologists' performance—initial experience. Radiology 2003; 227:496–474.
13. 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; 178:857–663.
14. 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.
15. Rubin GD, Lyo JK, Paik DS, et al. Pulmonary nodules on multi-detector row CT scans: performance comparison of radiologists and computer-aided detection. Radiology 2005; 234:274–283.
16. Li F, Arimura H, Suzuki K, et al. Computer-aided diagnosis for detection of missed peripheral lung cancers on CT: ROC and LROC analysis. Radiology 2005; 237:684–690.
17. 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.
18. Sone S, Takashima S, Li F, et al. Mass screening for lung cancer with mobile spiral computed tomography scanner. Lancet 1998; 351:1242–1245.
19. Li F, Sone S, Abe H, et al. Low-dose CT screening for lung cancer in a general population: characteristics of cancer in non-smokers versus heavy smokers. Acad Radiol 2003; 10:1013–1020.
20. 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.
21. 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.
22. Swensen SJ, Jett JR, Hartman TE, et al. Lung cancer screening with CT: Mayo Clinic experience. Radiology 2003; 226:756–761.
23. Pasmantier MW, Miettinen OS. CT screening for lung cancer: suspiciousness of nodules according to size on baseline scans. Radiology 2004; 231:164–168.
24. Henschke CI, Yankelevitz DF, Mirtcheva R, et al. CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules. AJR Am J Roentgenol 2002; 178:1053–1057.
25. Hasegawa M, Sone S, Takashima S, et al. Growth rate of small lung cancers detected on mass CT screening. Br J Radiol 2000; 73:930–937.
26. Benjamin MS, Drucker EA, McLoud TC, et al. Small pulmonary nodules: detection at chest CT and outcome. Radiology 2003; 226:489–493.
27. MacMahon H, Austin JHM, Gamsu G, et al. Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. Radiology 2005; 237:395–400.

Explanation of the Mechanism by which CAD Assistance Improves Diagnostic Performance when Reading CT Images

Toru Matsumoto*[a], Shinichi Wada[b], Shinji Yamamoto[c], Kohei Murao[d], Akira Furukawa[a], Masahiro Endo[a], Mitsuomi Matsumoto [e], Shusuke Sone[f]
[a]National Institute of Radiological Sciences; 9-1,Anagawa-4-chome, Inage-ku, Chiba-shi, 263-8555, JAPAN
[b]Niigata University;746 2-bancho, Asahimachi-dori, Niigata-shi, 951-8518, JAPAN
[c]Chukyo University, 101 Tokodachi, Kaizu-cho,Toyota-shi, 470-03936, JAPN
[d]IT Business Development Group, Fujitsu Ltd, 17-25, Shinkamata 1-chome, Ota-ku, Tokyo-to, 144-8588, JAPAN
[e] Diichi Hospital; 1227, Shimokotori-cho, Takasaki-shi, Gunma-ken, 370-0074, JAPAN
[f]JA Azumi General Hospital; 3207-1 Ikeda-ooaza, Ikeda-cho, Kitaazumi-gun, Nagano-ken, 951-8518,JAPAN

ABSTRACT

The purpose of our research is to make clear the mechanism that a reader (physician or radiological technologist) effectively identify abnormal findings in CT images of lung cancer screening by using with CAD system. A method guessing the 2X2 decision matrix between reader / CAD and reader / reader with CAD was investigated. We suppose the next scene to be it. At first, a reader judges whether abnormal findings per one patient per one CT image are present (1) or absent (0) without CAD results. The second, a reader judges whether abnormal findings are present (1) or absent (0) with CAD results. We expresses the correlation between diagnoses by a reader and CAD system for abnormal cases and for normal cases by following formula using phi correlation coefficient:$\varphi=(cd-ab)/\sqrt{(a+c)(b+d)(b+c)(a+d)}$. a,b,c,d: 2X2 decision matrix parameters. If $TPR1=(a+c)/n$, $TPR2=(b+c)/n$ and $TPR3=(a+b+c)/n$ for abnormal cases, $TPR3=TPR1+TPR2 - TPR1\times TRR2 - \varphi\sqrt{TPR1(1-TPR1)TPR2(1-TPR2)}$. Therefore, $a=n\ (TPR3 - TPR1)$, $b=n\ (TPR3 - TPR2)$, $c=n\ (TPR1 + TPR2 -TPR3)$, $d=n\ (1.0 - TPR3)$. This theory was applied for the experimental data. The 41 students interpreted the same CT images [no training]. A second interpretation was performed after they had been instructed on how to interpret CT images [training], and third was assisted by a virtual CAD [training + CAD]. The mechanism that makes up for a good point of a reader and a CAD with CAD in interpreting CT images was theoretically and experimentally investigated. We concluded that a method guessing the decision matrix (2X2) between a reader and a CAD decided the" presence" or "absence" of abnormal findings explain the improvement mechanism of diagnostic performance with CAD system.

Keywords: mechanism of double check, CAD, TNR, TPR, decision matrix, phi coefficient

1. INTRODUCTION

With advances in the development and technical performance of CAD, the need has arisen to study how CAD affects human readers, and the mechanism by which CAD improves diagnostic accuracy. Such research should lead to an elucidation of the circumstances in which the use of CAD will prove beneficial. At present, many of the research papers showing that CAD assistance improves diagnostic accuracy compared the accuracy of detection or differential diagnosis for a human reader working independently, with that obtained for a reader plus CAD output, by analysis of receiver operating characteristic (ROC) curves or by the area under the ROC curve (Az) [1-5].

To demonstrate whether or not CAD is actually useful in the clinical arena, we propose that CAD assistance should be evaluated by a single operating point (diagnosis of 0 or 1: binary protocol) on the ROC curve used in the clinical setting, rather than by the ROC curve overall. In brief, we propose a binary protocol in which the benefit of CAD is evaluated in accordance with the practice of informing the patient of the diagnostic result of 0 or 1, with reference by the reader to the CAD output as 0 or 1. To corroborate the validity of this protocol, we conducted a reading experiment for CT

matsu@nirs.go.jp

Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment, edited by Yulei Jiang, Miguel P. Eckstein, Proc. of SPIE Vol. 6146, 614619, (2006) · 1605-7422/06/$15 · doi: 10.1117/12.652928

images, and applying this protocol to the reading data, we evaluated CAD performance (complementarity) and reader performance (CAD consultation capacity, reading reproducibility). Via the above research process, we explain the mechanism by which CAD assistance improves diagnostic accuracy.

2. THEORETICAL DISCUSSION

We propose the following model to explain the mechanism by which diagnostic accuracy is improved by the use of CAD.

2.1 The protocol for CAD assistance studied in this research

The CAD protocol studied in this research is shown in Figure 1. First, a human reader and CAD independently read the same image database. Next, the reader reads each image with reference to the CAD output (hereafter, abbreviated as reader +CAD). The reader reads all images twice.

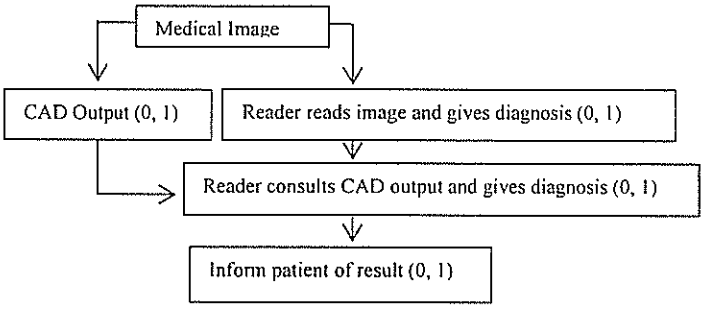

Medical Image
CAD Output (0, 1)
Reader reads image and gives diagnosis (0, 1)
Reader consults CAD output and gives diagnosis (0, 1)
Inform patient of result (0, 1)

Figure 1 Standard Protocol for CAD Use.
0: Detailed examination not required, 1: Detailed examination required, or follow-up (Not the same as " detailed examination not required") In the clinical setting, diagnoses of 0 and 1 by reader and CAD are fundamental.

2.2 Format of image diagnosis results in this report

The diagnoses produced by a reader, CAD, and reader +CAD are assumed to be responses of either 0 or 1 (normal or abnormal, or lesion absent or present) for the patient's CT image as a whole. For responses according to a continuous scale of percent-confidence level (0-100%), a level in the range 0-49% is classified as 0 and in the range 50-100% as 1. When the results are displayed in the ROC curve coordinate system (X-axis, FPR=1-TNR: 0-100%; Y-axis, TPR:0-100%), the diagnostic accuracies for reader, CAD, and reader +CAD are displayed as a pair of TNR and TPR values for each.

2.3 Effect of CAD assistance studied in this chapter

Figure 2 shows a typical example, displayed in the ROC curve coordinate system, of the results of diagnosis of 0 or 1 for reader, CAD, and reader +CAD for a certain image database. It shows that a reader with the performance represented by the point (TNR1, TPR1) on the ROC curve obtained the result represented by the point (TNR3, TPR3) on the ROC curve after consulting the output of the CAD system with performance represented by the point (TNR2, TPR2) on the ROC curve.

When the reader and CAD results were as shown in Figure 2, the reader +CAD results can be divided into any of the following three regions. First, that both the reader +CAD's TNR3 and TPR3 values are better than those of the reader and CAD. In the second region, one of the reader +CAD's TNR3 or TPR3 values is better than those of the reader or CAD, but the other was worse. In the third region, both the reader +CAD's TNR3 and TPR3 values are worse than those of either reader or CAD.

The results for reader +CAD in the first region show that CAD assistance was clearly worthwhile. The results for the third region show that there was clearly no value in the use of CAD. The results for the second region produce divergent evaluations. In brief, CAD assistance may be worthwhile when TNR3 (reader +CAD)>TNR1 (reader) or TPR3>TPR1. At this time, it was not possible to improve on the results for CAD alone, even after referring to the CAD output when TNR3<TNR2 or TPR3<TPR2. However, because the results for the reader alone were improved by consulting the CAD output, we can conclude that the use of CAD was worthwhile. In the opposite case, there was no value to using CAD. In brief, this refers to the situation when TNR3<TNR1 or TPR3<TPR1. Even if TNR3>TNR2 or TPR3>TPR2, the result for reader +CAD is worse than that obtained for the reader alone, and the CAD information is not useful for the reader. Further, in the situation TNR3=TNR1 or TPR3=TPR1 too, there is no value in using CAD. The fact that the same result was obtained with CAD assistance as for CAD alone may have significance for reconfirmation, but the time and effort required to read the same image twice raises doubts from the standpoint of labor saving.

In the present report, we separately discuss for normal and abnormal groups with definitive diagnoses, the framework (theory) in which the results TNR3>TNR1 and TPR3>TPR1 are produced in the first and second regions of Figure 2.

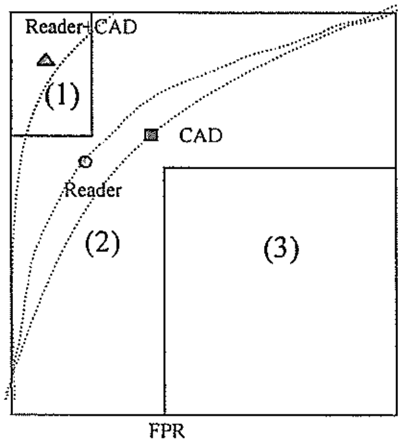


Figure 2 Three regions of the Reader +CAD results when the reader and CAD results were as shown in Figure

Diagnosis			Results	
Reader	CAD	R+C	Normal	Abnormal
1	1	1	FP	TP
		0	TN*	FN*
	0	1	FP*	TP
		0	TN	FN*
0	1	1	FP*	TP
		0	TN	FN*
	0	1	FP*	TP*
		0	TN	FN

Figure 3 Correspondence between the 0 and 1 diagnoses for reader, CAD, and reader +CAD. No asterisk: The case satisfied the criteria (i), (ii), (iii) for CAD consultation, Asterisk: "CAD-noncompliant responses".

2.4 Establishment of criteria for CAD consultation

The correspondence between the 0 and 1 diagnoses for reader, CAD, and reader +CAD—in accordance with the protocol for CAD use (Figure 1)—and the definitive diagnoses (true response)[5] is shown in Figure 3. Here, the criteria for CAD consultation to be satisfied by the reader for the 100% effective use of CAD output, via the opportunity of reading each image twice, while maintaining the consistency of diagnosis, are set out as follows.

(i)A true response by a reader leads to a true response for reader +CAD, regardless of the truth/falsity of the CAD response.

(ii)The reader response is false, but a true response for CAD leads to a true response for reader +CAD.

(iii)False responses by both reader and CAD also lead to a false response for reader +CAD.

The responses for reader +CAD can be divided into responses that satisfy the above three CAD consultation criteria (Figure 3, no symbol) and those that do not satisfy them (hereunder, denoted as a "CAD-noncompliant responses". Figure 3, asterisk). These are respectively the true responses (TN or TP) and false responses (FP or FN).

As shown in Figure 3 for example, in a situation when the reader diagnosis is 1 and the CAD diagnosis is also 1, a diagnosis of 1 by the reader +CAD is compliant with CAD consultation criterion (iii) in a normal individual, and is FP (false positive). In contrast, a diagnosis of 1 by the reader +CAD is compliant with criterion (i) in an abnormal individual, and is TP (true positive). If the reader diagnosis is 1, the CAD diagnosis is also 1, and the reader +CAD diagnosis is 0, then this diagnosis in a normal individual is noncompliant with criterion (iii) (noncompliant response:

based on inconsistency with the reader diagnosis), and the result is TN* (true response). In contrast, in an abnormal individual, it is noncompliant with criterion (i) (noncompliant response), and the result is FN* (false response). When the reader diagnosis is 1, the CAD diagnosis is 0, and the reader +CAD diagnosis is 1, this is a noncompliant response in which criterion (ii) has not been satisfied in a normal individual, and is FP* (false response). In contrast, the same pattern for an abnormal individual produces a TP (true response). Further, when the reader diagnosis is 0, the CAD diagnosis is 1, and the reader +CAD diagnosis is 1, the result in an abnormal individual is a FP* (false response), as a consequence of noncompliance with criterion (i). In contrast, the same pattern in which criterion (ii) is satisfied in an abnormal individual produces a TP (true response).

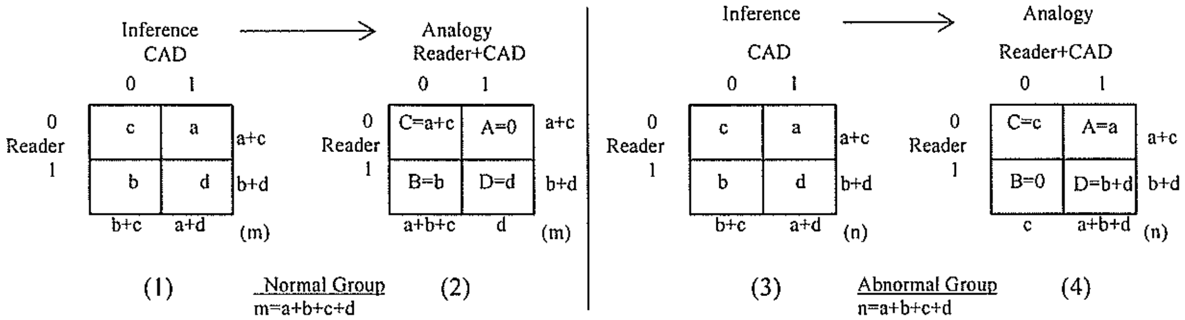


Figure 4 Estimation of 4-way contingency table for reader/CAD to achieve the reader +CAD

2.5 Four-way contingency tables based on diagnoses of 0 and 1 by reader/CAD and reader/reader +CAD

Figure 4 shows the results obtained when the same image database is independently read by a reader and CAD, and a 4-way (2 x 2) contingency table (hereunder, abbreviated to "contingency table") is displayed for the differences between reader and CAD diagnoses for the normal group and abnormal group,. Figure 4 (1) and (2) show contingency tables for the normal group and Figure 4 (3) and (4) for the abnormal group (NB: hereunder in this text, the contingency table parameters are similarly displayed as a, b, c, and d. Note that a, b, c, and d>0.)

Figure 4 (2) shows a contingency table (A, B, C, D>0) for the normal group and abnormal group displaying the differences between diagnoses of the reader and reader +CAD in situations when the CAD consultation criteria (i), (ii), and (iii) are completely satisfied by the reader (Figure 3 CAD-noncompliant response 0: in all asterisked cases), for the normal group, with Figure 4 (2). In brief, in accordance with CAD consultation criterion (i), A=0, C=a+c. In accordance with CAD consultation criterion (ii), B=b and in accordance with CAD consultation criterion (iii), D=d. In contrast, Figure 4 (4) shows the same concept for the abnormal group. In brief, in accordance with criterion (i), B=0 and D=b+d, and in accordance with criterion (ii), A=a, and in accordance with criterion (iii), C=c.

2.6 Estimates of contingency tables for reader/CAD and reader/reader +CAD to achieve accuracy of reader+CAD (observed TNR3 and TPR3 respectively) (when the CAD consultation criteria are completely satisfied)

Via the protocol for the use of the CAD in Figure 1, with known (TNR1, TNR2, TNR3) and (TPR1, TPR2, TPR3) for an arbitrary reader, CAD, and reader +CAD located in the ROC curve coordinate system, as shown in Figure 2, a method for estimating a, b, c, and d of Figure 4 (1) and (3) for the contingency table for reader/CAD to produce the TNR3 and TPR3 values for reader +CAD is shown below. Based on those results, a method for inferring by analogy the A, B, C, and D values for Figure 4 (2) and (4), the contingency table for reader/reader+CAD, is shown. This assumes, however, that the CAD consultation criteria (i), (ii) and (iii) are completely satisfied.

2.6.1 For the normal group (number of images: m)

When the reader TNR1=(a+c)/m, CAD TNR2=(b+c)/m, and the reader satisfies consultation criteria (i), (ii), and (iii) (CAD-noncompliant responses for Figure 3, FP*=0, TN*=0), the TNR3 for reader +CAD is shown as TPR3=(a+b+c)/m. The contingency table parameters for reader and CAD are thus a=m (TNR3–TNR2), b=m (TNR3–

TNR1), c=m (TNR1+TNR2–TNR3), and d=m (1.0–TNR3). Hence, the contingency table parameters for reader and reader +CAD, are, in accordance with Figure 4 (2), A=0, B=b=m (TNR3–TNR1), C=a+c=m (TNR1), D=d=m (1.0–TNR3). Provided however, that TNR3≥TNR2 or TNR1.

2.6.2 For the abnormal group (number of images: n)

When the reader TPR1=(b+d)/n, CAD TPR2=(a+d)/n, and the reader satisfies consultation criteria (i), (ii), and (iii) (CAD-noncompliant responses for Figure 3, FN*=0, TP*=0), the TPR3 for reader +CAD is shown as TPR3=(a+b+d)/n. The contingency table parameters for reader and CAD are thus a=n (TPR3–TPR1), b=n (TPR3–TPR2), c=n (1.0–TPR3), and d=n (TPR1+TPR2–TPR3). Hence, the contingency table parameters for reader and reader +CAD, are, in accordance with Figure 4 (4), A=a=n (TPR3–TPR1), B=0, C=c=n (1.0–TPR3), and D=b+d=n (TPR1). Provided however, that TPR3≥TPR2 or TPR1.

3. EMPIRICAL DISCUSSION

We conducted a reading experiment to compare with actual practice, the theory behind the mechanism by which diagnostic accuracy is improved by the protocol for CAD assistance in Figure 1, and analyzed the results by the method described below.

3.1 Reading experiment[6]

Forty-one medical technology students who could become CT screeners (CTS) in the future participated in the reading experiment described below. In the first experiment, the readers participated in a training session for reading CT images consisted of single-slice image. Next, they read another single-slice CT images for a total of 99 patients—73 abnormal and 26 normal individuals—with definitive diagnoses, and evaluated the presence/absence of an abnormal finding. If a reader was detected a abnormal finding, the lesion was localized on a chest CT diagram by a reader, and a response was made according to a continuous confidence scale of 0-100%. One month later, the second experiment was performed. Between the first and second experiments, the CTS attended four, once-weekly lectures on the CT diagnosis of chest diseases. Immediately before the second experiment, the CTS underwent practice exercises for a separate CT image database from that used in the first experiment. Next, the CTS read the same images as in the first experiment, and evaluated them for the presence/absence of an abnormal finding. Three months later, the CTS consulted the CAD results for the CT images for the same patients, and evaluated the presence/absence of an abnormal finding by the same method as in the first and second experiments. At this time, the results of the CAD presented to CTS were hypothetical, prepared artificially by the researchers to have the average performance of the CTS in the second experiment. There was only one abnormal finding detected by the hypothetical CAD, and its location was circled on the CT image. If an abnormal finding by the hypothetical CAD was not detected, the notation "No abnormal finding" was made on the CT image. The CTS were informed before the reading experiment that the performance of the hypothetical CAD was FPR=8.3% (TNR=91.7%) and TPR=80% (FNR=20%).

Throughout the first, second, and third experiments, the readers were instructed to detect only the most important abnormal finding.

3.2 Classification of 0 and 1 diagnoses for experimental data (TP, FN, FP, TN)

Based on whether an abnormal finding was present (1) or absent (0) on each single-slice CT image, the CTS response was also either 1 or 0. However, CTS also added a continuous confidence scale to their diagnosis.

An image with a definitive diagnosis of "abnormal finding present" was considered to be composed of a finding-present region and a finding-absent region, and a definitive diagnosis of "abnormal finding absent" was considered to be composed only of a finding-absent region. This process yielded a total of 73 finding-present regions and 99 finding-absent regions among 99 cases read. For a case with a definitive diagnosis of abnormal finding present, once a decision was made on a diagnosis of either a finding-present region or a finding-absent region (presence/absence of a finding plus a confidence scale (%)), the diagnosis for the other region is automatically also decided. A confidence level of 0-49% was classified as 0 and 51-100% as 1 (an evaluation of 50% was forbidden). For a finding-present region, the CTS diagnosis was classified as TP when an abnormal finding was made at the same location as the definitive diagnosis, and as FN when it was overlooked. For a finding-absent region, the CTS diagnosis was classified as FP when an abnormal finding was detected, and as TN when a finding was not detected. Furthermore, the diagnostic results of the hypothetical CAD were classified as TP, FN, FP, or TN in the same manner as for the CTS. The results obtained when the hypothetical CAD was consulted (reader +CAD) were similarly classified (see the references for further details).

3.3 Method of analysis of experimental data

3.3.1 Comparison of measured data and estimates

Using the 0 and 1 diagnosis data for reader, CAD, and reader +CAD, the TNR and TPR values were calculated for normal and abnormal individuals, and based on those results, contingency tables were estimated for reader/CAD, and reader/reader +CAD when the CAD consultation criteria (i), (ii), and (iii) were completely satisfied. Additionally, a contingency table for reader/reader +CAD was determined from the measured data for reader and reader +CAD (including the CAD-noncompliant responses). Next, the deviations between the estimates and the measured values were determined for the same contingency table parameters as above. In short, Deviation=|Estimated contingency table parameters–Measured contingency table parameters|. Moreover, the CAD-noncompliant responses included in all responses A, B, C, and D of the contingency table parameters for reader and reader +CAD were identified and tabulated.

3.3.2 Evaluation of reader and CAD complementarity[7]

There are two types of contingency table for reader/CAD. One is the contingency table for measurements that can be prepared from the reader's 0 and 1 diagnoses and the CAD's 0 and 1 diagnoses. The other is the estimated contingency table for reader/CAD that produces the measured TNR3 or TPR3 for reader +CAD. From these contingency table data a, b, c, and d for reader/CAD, the complementarity of reader and CAD was evaluated by means of the φ coefficient.

$\varphi=(cd-ab)/\sqrt{(a+c)(b+d)(b+c)(a+d)}$, a, b, c, and d>0.

The TNR or TPR for reader, CAD, and reader +CAD were respectively labeled X1, X2, and X3, and the complementarity φ of reader and CAD was evaluated according to the following equation.

$\varphi=(X1+X2-X1\cdot X2-X3)/\sqrt{X1(1-X1)X2(1-X2)}$

Where $-1\leq\varphi\leq+1$. This implies that when φ is -1, the complementarity is a maximum (similarity is a minimum), when φ=0, they are mutually independent, and when φ=+1, the complementarity is a minimum (the similarity is a maximum). When TNR3<TNR1 or TNR2, and when TPR3<TPR1 or TPR2, a or b will be negative, in which case φ will be meaningless.

3.3.3 Evaluation of CAD assistance effect

From the contingency tables, for each of the estimates (complete satisfaction of the CAD consultation criteria) and the measured values (including CAD-noncompliant responses), the following p-values (one-sided) were calculated, and the respective CAD assistance effects were evaluated.

$$\text{Exact } p=(1/2)^{a+b}\sum_{j=0}^{r}{}_{a+b}C_j\text{, where } r=\min(a,b)$$

4. EXPERIMENTAL RESULTS

4.1 Example of estimation of contingency tables for reader/CAD, and reader/reader+CAD to produce measured reader +CAD diagnostic accuracy (TNR3, TPR3)

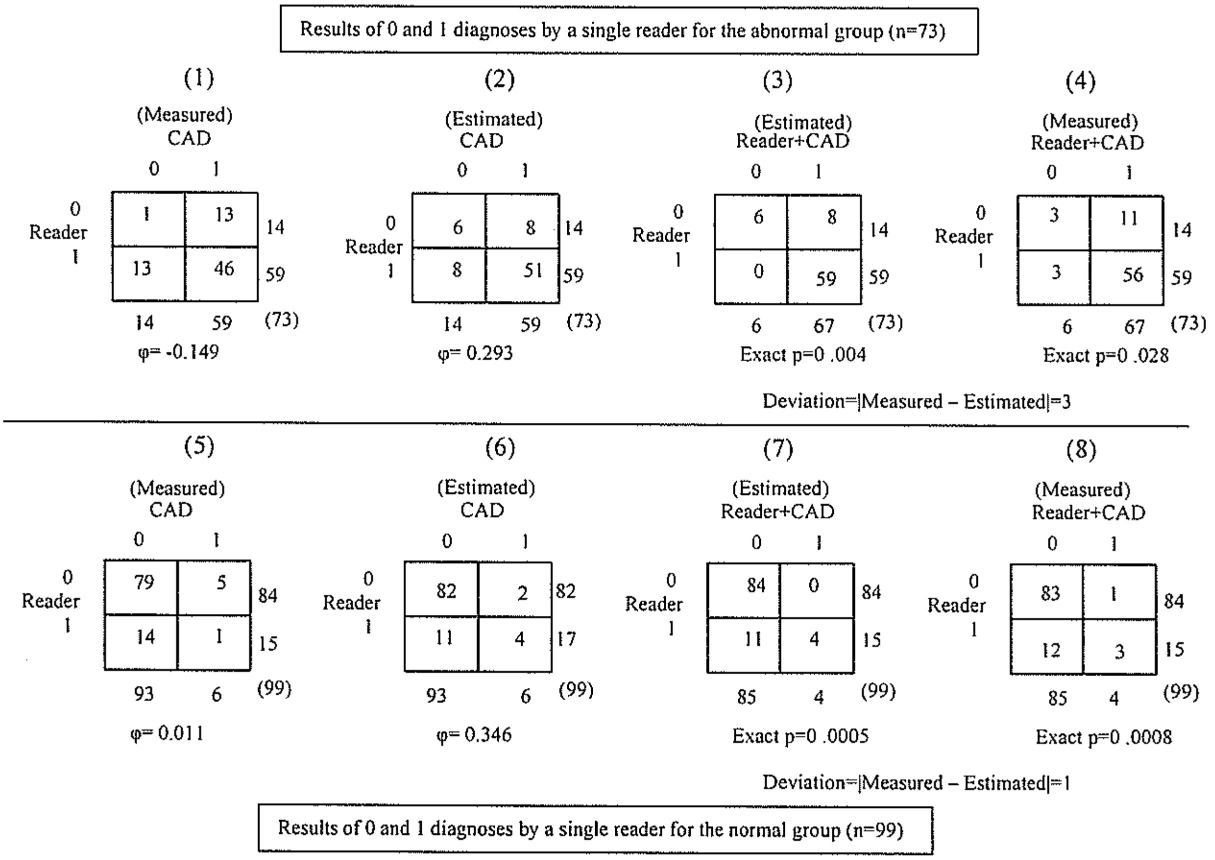
4.1.1 Results of 0 and 1 diagnoses by a single reader for the abnormal group (n=73) (Figure 5):

The number of true responses (TP) for the reader alone was 59, and the number of TP for CAD was also 59. The measured contingency table for reader/CAD at this time is shown in Figure 5 (1). If this complementarity can be completely utilized by a reader, the expected number of TP for reader +CAD would be 72=13+13+46, yielding a complementarity in this case of φ=-0.149. The number of TP for reader +CAD obtained in actual practice was

67=11+56, determined from the contingency table for reader/reader +CAD (Figure 5 (4)). The estimated contingency table for reader/CAD to produce this was Figure 5 (2). The number of TP for reader +CAD in this case was 67=8+8+51. The complementarity of reader and CAD at this time has decreased to (NB: complementarity is higher at lower values of φ). From the estimated contingency table, the contingency table for reader/reader +CAD when the reader has satisfied the CAD consultation criteria was estimated to be Figure 5 (3). The p-value for CAD assistance effect at this time (difference between reader and reader +CAD) was p=0.004. In contrast, the p-value for the measured reader and reader +CAD was p=0.028. In the contingency table for reader/reader +CAD (Figure 5 (4)), the CAD-noncompliant response was B for 3 cases (1 diagnosis for the reader: true response, and 0 diagnosis for CAD: false response, and 0 diagnosis for reader +CAD: false response), and C for 2 cases (0 diagnosis for reader: false response, and 1 diagnosis for CAD: true response, and 0 diagnosis for reader +CAD: false response), for a total number of 5 false responses. The deviations (absolute value) in the contingency table parameters for the measured (Figure 5 (4)) and estimated (Figure 5 (3)) reader and reader +CAD were equal for each parameter in 3 cases. In brief, there were 3 more cases in which a finding was overlooked by the reader alone and detected by reader +CAD (A), and 3 cases (B) were overlooked by reader +CAD despite detection by the reader alone. The estimated contingency table when these deviations were offset was obtained (Figure 5 (3)).

4.1.2 The results obtained for the normal group (m=99)

Figure 5 (5) to (8) are shown that similar tendencies were displayed as those in the abnormal group.



Results of 0 and 1 diagnoses by a single reader for the abnormal group (n=73)

(1) (Measured) CAD

		0	1	
Reader	0	1	13	14
	1	13	46	59
		14	59	(73)

φ= -0.149

(2) (Estimated) CAD

		0	1	
Reader	0	6	8	14
	1	8	51	59
		14	59	(73)

φ= 0.293

(3) (Estimated) Reader+CAD

		0	1	
Reader	0	6	8	14
	1	0	59	59
		6	67	(73)

Exact p=0 .004

(4) (Measured) Reader+CAD

		0	1	
Reader	0	3	11	14
	1	3	56	59
		6	67	(73)

Exact p=0 .028

Deviation=|Measured – Estimated|=3

(5) (Measured) CAD

		0	1	
Reader	0	79	5	84
	1	14	1	15
		93	6	(99)

φ= 0.011

(6) (Estimated) CAD

		0	1	
Reader	0	82	2	82
	1	11	4	17
		93	6	(99)

φ= 0.346

(7) (Estimated) Reader+CAD

		0	1	
Reader	0	84	0	84
	1	11	4	15
		85	4	(99)

Exact p=0 .0005

(8) (Measured) Reader+CAD

		0	1	
Reader	0	83	1	84
	1	12	3	15
		85	4	(99)

Exact p=0 .0008

Deviation=|Measured – Estimated|=1

Results of 0 and 1 diagnoses by a single reader for the normal group (n=99)

Fig.5 Example of estimation of contingency tables for reader/CAD, and reader/reader+CAD to produce measured reader +CAD diagnostic accuracy for abnormal (upper), normal (lower)

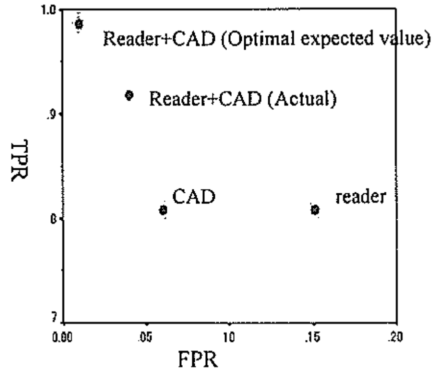


Figure 6 Results of diagnostic accuracy for reader, CAD, and reader +CAD in the ROC curve coordinate system

4.1.3 Representation of diagnostic accuracy for reader, CAD, and reader +CAD in the ROC curve coordinate system

Figure 6 shows the outcome when above reading results are displayed in the ROC curve coordinate system. The actual diagnostic accuracy was lower than the reader +CAD figure expected from the original observed values.

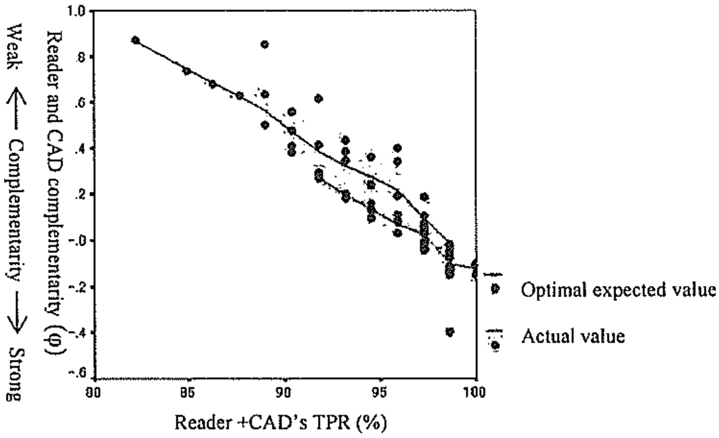


Figure 7 Relationship between reader +CAD's TPR and the complementarity (φ)

4.2 Diagnostic accuracy of reader +CAD as a function of the reader and CAD complementarity φ

Figure 7 shows the relationship between the reader/CAD complementarity φ and the corresponding measured TPR3 values for reader +CAD when 41 readers read 73 abnormal cases according to the protocol for CAD use in Figure 1. There are two types of relationship at work here. One is the relationship between the φ calculated from the measured contingency table (Figure 5 (1)) and the expected diagnostic accuracy (TPR3) when the reader/CAD complementarity is 100% utilized, and the other is the relationship between the actual diagnostic accuracy (TPR3) and the φ calculated from the estimated contingency table (Figure 5 (2)) to produce this accuracy. Both relationships show a tendency for the diagnostic accuracy to be higher when CAD consultation is associated with greater complementarity.

4.3 The |deviation| between the estimated and observed values for A, B, C, and D of the contingency table for reader/reader +CAD

The deviations between the values of A, B, C, and D in the estimated reader/reader +CAD contingency table and the measured values was determined from the diagnoses of 0 and 1 obtained from 41 readers reading 99 images. This process showed that the deviations were equivalent for each parameter for all readers producing deviations for the 73 cases in the abnormal group. Among the 41 readers, the deviation was 0 for 6 readers (15%), 1 for 10 (24%), 2 for 7 (17%), 3 for 3 (7.3%), 4 for 4 (10%) 5 for 4 (10%), and other for 7 readers (17%).

4.4 Diagnostic accuracy for reader +CAD as a function of the number of CAD-noncompliant responses

Figure 8 shows the correspondence between results of tabulation of the total number of CAD-noncompliant responses by each reader included in the responses for the measured contingency table parameters A, B, C, and D for reader/reader +CAD, and the diagnostic accuracy (TPR3, TNR3: measured values) for each reader. This shows that fewer noncompliant responses is associated with a higher tendency for diagnostic accuracy.

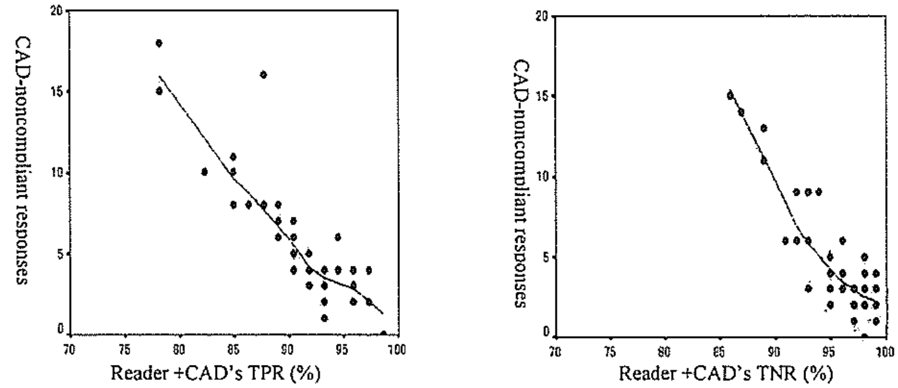


Figure 8 Relationship between CAD-noncompliant responses and the reader +CAD's TPR and TNR(%): CAD-noncompliant responses is , 1) Reader True and CAD False result in Reader +CAD False, 2) Reader False and CAD True result in Reader +CAD False, 3) Reader False and CAD False result in Reader +CAD True

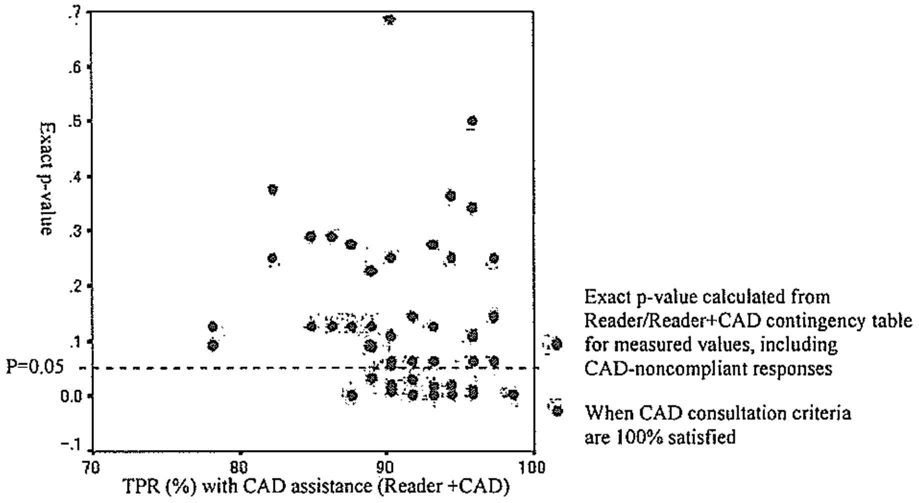


Figure 9 Test of significant difference between reader and reader +CAD's TPR values

4.5 Test of significant difference in the effect of CAD assistance

Figure 9 shows the relationship between the exact p-value—that represents the degree of difference in diagnostic accuracy (TPR) for reader/reader +CAD—and the actual measured diagnostic accuracy (TPR3). The two do not appear to be correlated. As regards the difference in diagnostic accuracy (TPR) between reader and reader +CAD, there were more readers yielding statistical significant differences ($p<0.05$) when the CAD consultation criteria were satisfied, in comparison with the actual situation (including CAD-noncompliant responses).

5. DISCCUSION

The development of CAD is presently flourishing. Several systems has been approved by the FDA, and are marketed as a medical device and used in the clinical setting. However, we are yet to see adequate investigations into the effect of CAD on human readers and how they are actually useful in practice. In the present work, we sought to elucidate the mechanism by which CAD yields a benefit, on the assumption of the situation shown in Figure 2, when CT images are given diagnoses of 0 or 1 after the reader consults the CAD output. Here, to evaluate first whether or not CAD is useful in practice, we note that a method in which the diagnosis is 0 and 1 is more realistic than using the ROC curve overall. Next, we discuss the significance of the contingency table estimation method for reader/CAD, and reader/reader +CAD proposed in this report. Moreover, by analyzing the deviation between the estimates and actual measured values, we are able to explain the mechanism by which the CAD assistance is obtained.

5.1 Necessity for a binary evaluation method

5.1.1 The foundation of the image diagnosis is the diagnosis of 0 or 1.

In the protocol for CAD use method shown in Figure 1, we consider the situation when a reader will use CAD in actual practice. In this report, the CAD result presented to the reader is assumed to be a diagnosis of 0 or 1 for each patient. Such a CAD protocol is applicable to a CAD for differential diagnosis of normal or abnormal (benign or malignant) (Reference). The CAD performance at that time is denoted by one point on the ROC curve. In contrast, for CAD intended to detect abnormal findings, it is normal for CAD to produce numerous FPs in an image for one patient. Hence, when the result of a diagnosis of 0 or 1 for each patient is presented to the reader, the CAD's TPR2 fits within the range 0-100%, but it is highly likely that the FPR will be 100%. Even in such circumstances however, the method described in Section 2.6 of this report will be effective. The estimation of the contingency tables for reader/CAD, and reader/reader +CAD can be performed for the TNR or TPR for three parties—an arbitrary reader, CAD, or reader +CAD—within the ROC curve coordinate system (including data on the coordinate axes).

In contrast, a reader diagnoses the same image twice, once as a reader alone and again with CAD assistance. In each of these situations, regardless of the type of diagnosis—detection (detection of abnormal findings), differential diagnosis, etc.—the reader's diagnosis of 0 or 1 is the basis. A slightly more detailed diagnosis than the 0 or 1 result is a response in terms of a confidence level between 0 and 1. The reader's diagnostic accuracy will be evaluated later by ROC curve, but the ROC curve itself is not communicated to the patient. The patient will be notified of the diagnosis of either 0 (for example, no abnormality), or not 0 (1: for example, abnormality present or follow-up). When a single reader has diagnosed the medical images of numerous patients according to a confidence level between 0 and 1, the diagnostic performance of that reader is represented by a single ROC curve, but the result of 0 or 1 diagnosed by that reader is represented by a single point on this ROC curve.

Summing up the above, when a reader in the diagnostic setting with the performance (TNR1, TPR1) represented by one point on the ROC curve consults the CAD output (TNR2, TPR2) represented by one point on the ROC curve, a result (TNR3, TPR3) for CAD consultation (reader +CAD) represented by one point on the ROC curve is produced.

5.1.2 The horizontal axis (FPR=1−TNR) of imaging diagnosis undertaken in the clinic is approximately constant.

Diagnosis by a reader alone is performed so that the horizontal axis (FPR=1−TNR) of the ROC curve is constant. No other horizontal axis (FPR1=1−TNR1) is used. In Japan, the screening examination for various cancers is undertaken as an arm of government policy, but the rate of detailed examination required produced by primary screening (approximately equivalent to the FPR) is accuracy-controlled to be constant. In the case of lung cancer screening using chest X-ray films for example, the mean FPR is 2.6% of all readings. Hence, the result of detecting the difference in the diagnostic accuracy between "without CAD (reader)" and "with CAD (reader +CAD)" using the ROC curve overall (Az) does not necessarily guarantee that CAD will be useful in the clinical arena. A method for detecting a difference in the TPR for reader/reader +CAD when the horizontal axis (FPR=1–TNR) is constant may be the CAD evaluation method that corresponds to actual reality.

5.2 Significance of the estimation of the contingency tables for reader/CAD and reader/reader+CAD

As shown in the example in Figures 5 and 6, if there are results of diagnoses of 0 and 1 for the same database by the reader and CAD, a contingency table for reader/CAD is obtained. And, the optimal outcome can be expected if a reader fully utilizes the complementarity between reader and CAD shown in the contingency table. However, this is a numeral showing a possibility, and the actual reality is not necessarily the same. As shown in Figure 6, the accuracy of reader +CAD in practice is either the same as, or normally below the optimal expected value. This is because the reader has two opportunities for reading an image and making a CAD-noncompliant response such as those marked with asterisks in Figure 3. And, it is not possible to identify the outcome for the contingency table for reader/CAD at that time from the data only from diagnoses of 0 and 1 made independently by the reader and CAD. In this report, the method for estimating the contingency table for reader/CAD to produce the result for reader +CAD in actual practice is shown in 2.6.

The actual reader/CAD complementarity φ in Figure 7 was lower than the complementarity of reader and CAD to produce the optimal expected accuracy of reader +CAD (TNR3 or TPR3). It can also be seen from Figure 8 that the greater the complementarity between reader and CAD, the higher the diagnostic accuracy that can be obtained. These facts lead to the suggestion that for CAD to be useful to readers, it will be necessary for CAD to output numerous instances of true responses, despite false responses by readers.

The contingency table for reader/reader +CAD can be inferred by analogy from the estimated contingency table for reader/CAD. This estimated contingency table shows the situation when CAD consultation criteria (i), (ii), and (iii) described in chapter 2.4 are completely satisfied. The deviation between these estimated values and the observed contingency table parameters for reader/reader +CAD represents the difference from a reader using a CAD-noncompliant response. As shown in Figure 6 (an abnormal case), the absolute value of this deviation is equally produced in parameters A, B, C, and D of the contingency table. This phenomenon was observed for all readers producing deviations, not only for the abnormal group, but also for the normal group. Briefly, given the opportunity for the second reading (for reader +CAD), there were equal numbers of occurrences of false response for the reader alone but a true response after CAD consultation, and of a true response for reader alone and a false response after CAD consultation. The result of offsetting these true responses and false responses was resolved in the estimated contingency table for reader/reader +CAD. This fact corroborates the validity of the contingency table for reader/reader +CAD estimated using this method.

Among the 41 readers, the deviation between the estimate and observation was 0 for 6 readers (15%). For these cases, close investigation of the nature of the contingency table parameters A, B, C, and D for reader/reader +CAD, revealed the inclusion of CAD-noncompliant responses in the A, B, C, and D responses for the contingency table, as for Figure 5 (the situation for a deviation of 3). This signifies that a deviation of 0 between the estimation of contingency table parameters for reader/reader +CAD and the actual measurements that it was a apparent phenomenon (including CAD-noncompliant responses).

When we closely investigate the nature of the contingency table parameters A, B, and C for reader/reader +CAD, it is possible to identify which of the CAD consultation criteria (i), (ii), or (iii) has not been satisfied (one of those marked with an asterisk in Figure 3).

As shown in Figure 8, a higher diagnostic accuracy was yielded for readers with fewer noncompliant responses. As well, as shown in Figure 9, it was easy to produce a statistically significant difference in diagnostic accuracy between "without CAD (reader)" and "with CAD (reader +CAD)" when the CAD consultation criteria were satisfied, than when they were not satisfied. These facts suggest that the extent to which a reader can definitely satisfy the CAD consultation criteria influences the effective use of CAD. The fundamental challenge for the use of CAD lies in whether the correct diagnosis can be selected when the reader's diagnosis and the CAD output are inconsistent. In order to reliably achieve consistency, it is necessary, at least before using CAD, to thoroughly ascertain one's compatibility with CAD, focusing on the situations in which complementarity with CAD is present, and the extent of such complementarity.

5.3 Application of this method to other protocols for CAD use

Figure 10 presents an overview of protocols for CAD use other than that in Figure 1 that are conceivable at this point in time. Figure 10 (1) is the same as Figure 1, the protocol investigated in this report. It has been redrawn to shed light on differences with other protocols. In the following we designate it the standard protocol. In the standard protocol, the reader reads all images twice. By contrast, in the protocol in Figure 10 (2), the reader reads the image and makes an evaluation of "abnormality present" (a diagnosis of 1) without reference to the CAD output, and only consults the CAD output in cases where the evaluation is "abnormality absent" (a diagnosis of 0). An image in which the reader has made a diagnosis of 1 is read only once and one in which a diagnosis of 0 is made is read twice (reader/CAD contingency table parameters a and c). Figure 10 (3) represents a modification of the standard protocol. Only in the case where the reader diagnosis and CAD output are inconsistent (a, b) does the reader consult the CAD output. Figure 10 (4) represents a modification of the method in Figure 10 (2). First, if the reader makes a diagnosis of 1, a reader +CAD diagnosis is not made. Next, if the reader makes a diagnosis of 0 and the CAD output is also 0, a combined diagnosis is not made. CAD is consulted only in cases where a diagnosis of 1 is made (inconsistency between reader and CAD: a). Figure 10 (5) represents a method in which the reader only reads the image in cases where the CAD output is 0. In Figure 10 (6), all images are read once while the reader consults the CAD output.

From the reading experiment data investigated for the standard protocol shown in Figure 1 and Figure 10 (1), it is possible to simulate the reader and CAD complementarity, the number of CAD-noncompliant responses, and the reader +CAD diagnostic accuracy (TNR3, TPR3) obtained by the particular protocol for CAD use when the protocols described in Figure 10 (2) through (5) are employed. Here, we first discuss the advantages and disadvantages of the standard protocol and Figure 10 (2), and mention the characteristics of the other protocols.

In a situation when the CAD usage criteria are completely satisfied in the protocol of Figure 10 (2), TNR3 (=1-FPR3) will be the same as the TNR1 obtained for reader alone (TNR3=TNR1). It can be expected that TPR will benefit more than that with the standard protocol. The situation in this instance is shown in Figure 11 (1). That is to say, when such a situation has ensued, it would be suggested that the reader has made a diagnosis using the protocol of Figure 10 (2).

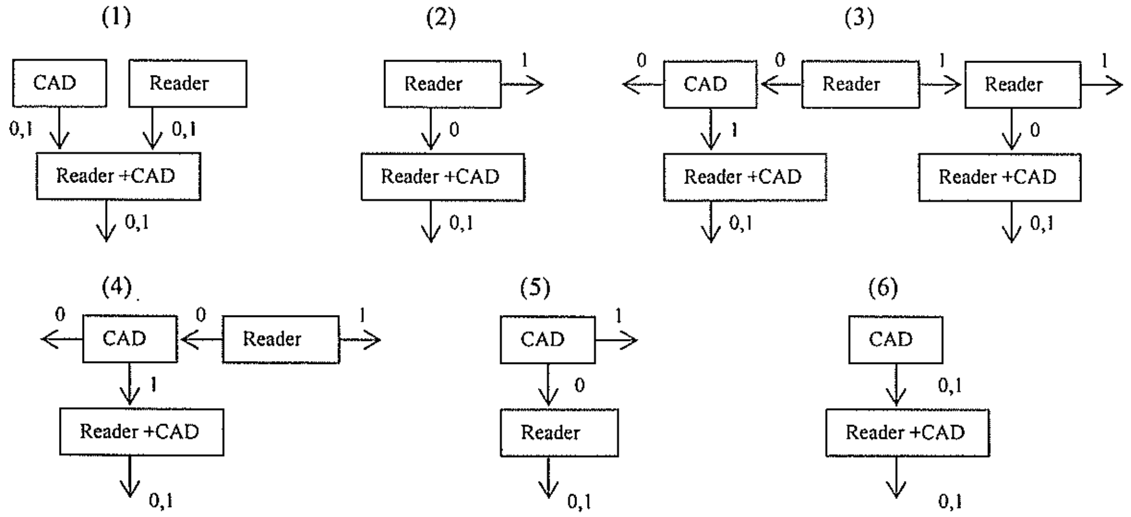


Figure 10 Overview of protocols for CAD use