Bin Zheng, PhD
Marie A. Ganott, MD
Cynthia A. Britton, MD
Christiane M. Hakim, MD
Lara A. Hardesty, MD
Thomas S. Chang, MD
Howard E. Rockette, PhD
David Gur, ScD

Index terms:
Breast neoplasms, diagnosis, 00.30, 00.81
Cancer screening, 00.11
Computers, diagnostic aid
Diagnostic radiology, observer performance

Published online before print
10.1148/radiol.2213010308
Radiology 2001; 221:633–640

Abbreviations:
A_z = area under the receiver operating characteristic curve
CAD = computer-assisted detection

[1] From the Division of Imaging Research, Department of Radiology (B.Z., D.G.), the Departments of Radiology (C.A.B., M.A.G., C.M.H., L.A.H., T.S.C.) and Biostatistics (H.E.R.), University of Pittsburgh, 300 Halket St, Suite 4200, Pittsburgh, PA 15213; and the Magee Womens Hospital, University of Pittsburgh Medical Center Health System, Pa (M.A.G., C.M.H., L.A.H.). Received January 12, 2001; revision requested March 5; revision received March 29; accepted May 1. Supported in part by the U.S. Army Medical Research Acquisition Activity under contracts DAMD17-98-1-8018 and DAMD17-00-1-0410 and by grant CA77850 from the National Cancer Institute, National Institutes of Health. **Address correspondence to** B.Z. (e-mail: *bzheng@radserv.arad.upmc.edu*).

The content of the contained information does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

© RSNA, 2001

See also the editorial by D'Orsi (pp 585–586) in this issue.

Author contributions:
Guarantors of integrity of entire study, B.Z., D.G.; study concepts and design, B.Z., D.G.; literature research, B.Z.; experimental studies, L.A.H., M.A.G.; data acquisition, B.Z.; data analysis/interpretation, B.Z., D.G., H.E.R.; statistical analysis, B.Z., H.E.R.; manuscript preparation, M.A.G., L.A.H.; manuscript definition of intellectual content, B.Z., D.G.; manuscript editing, T.S.C., M.A.G.; manuscript revision/review, C.M.H., C.A.B., D.G., B.Z., H.E.R.; manuscript final version approval, B.Z., D.G., H.E.R.

Soft-Copy Mammographic Readings with Different Computer-assisted Detection Cuing Environments: Preliminary Findings[1]

PURPOSE: To assess the performance of radiologists in the detection of masses and microcalcification clusters on digitized mammograms by using different computer-assisted detection (CAD) cuing environments.

MATERIALS AND METHODS: Two hundred nine digitized mammograms depicting 57 verified masses and 38 microcalcification clusters in 85 positive and 35 negative cases were interpreted independently by seven radiologists using five display modes. Except for the first mode, for which no CAD results were provided, suspicious regions identified with a CAD scheme were cued in all the other modes by using a combination of two cuing sensitivities (90% and 50%) and two false-positive rates (0.5 and 2.0 per image). A receiver operating characteristic study was performed by using soft-copy images.

RESULTS: CAD cuing at 90% sensitivity and a rate of 0.5 false-positive region per image improved observer performance levels significantly ($P < .01$). As accuracy of CAD cuing decreased so did observer performances ($P < .01$). Cuing specificity affected mass detection more significantly, while cuing sensitivity affected detection of microcalcification clusters more significantly ($P < .01$). Reduction of cuing sensitivity and specificity significantly increased false-negative rates in noncued areas ($P < .05$). Trends were consistent for all observers.

CONCLUSION: CAD systems have the potential to significantly improve diagnostic performance in mammography. However, poorly performing schemes could adversely affect observer performance in both cued and noncued areas.

Breast cancer is one of the leading causes of death in women over the age of 40 years (1,2). To reduce mortality and morbidity with early diagnosis and treatment, current guidelines recommend periodic mammography screening for women aged 40 and over (3). Due to the large number of mammographies performed and the low yield of abnormalities detected in screening environments, detecting abnormalities (mainly masses and microcalcification clusters) from the background of a complex normal anatomy is a tedious, difficult, and time-consuming task for most radiologists (4,5).

Hence, there is a growing interest in the development of computer-assisted detection (CAD) schemes for mammography. It is generally believed that such schemes could eventually provide radiologists with a valuable "second opinion" and help improve accuracy and efficiency of breast cancer detection at an early stage (6,7).

To assess the potential for improving diagnostic accuracy and efficiency in mammography, several studies have been performed by using the CAD systems. These studies have demonstrated that with the appropriate assistance of CAD systems, radiologists could either detect more subtle cancers in a screening environment (8,9) or increase the accuracy of distinguishing malignant lesions from those that are benign (10–12). While some authors (13–15) indicated that CAD did not substantially decrease the specificity levels of the radiologists, others (16,17) indicated that current CAD systems could significantly decrease diagnostic accuracy and efficiency of radiologists due to high false-positive

detection rates. As there is difficulty in comparing the performance of different CAD schemes developed at various institutions (18), the results of these studies are not easily comparable, since different CAD schemes, radiologists, and cases were included. Authors of these studies did not address in detail how CAD could affect the diagnostic performance of the observers or the level of CAD that may be required to be widely acceptable as a helpful tool in the clinical environment.

Researchers have suggested that large-scale experiments are needed to assess the effect of CAD (eg, the false-positive identifications) on the diagnostic accuracy of radiologists (19). Some doubt remains as to whether CAD systems might increase the number of unnecessary follow-up examinations or biopsies and thereby offset the benefits from the potential gains in sensitivity (20).

The effect of precuing images (highlighting suspicious areas) has been of great interest in the field of perception psychology in general (21,22) and of diagnostic radiology in particular (23–25). Much of the work was associated with attempts to improve tumor detection on x-ray images of the chest. In a series of carefully designed experiments, Krupinski et al (26) demonstrated that in a cued environment, performance of radiologists in detecting true-positive lung nodules that had not been cued was degraded substantially. The shapes of abnormalities (ie, masses and microcalcification clusters) and the complexity of the background tissue seen on mammograms are somewhat different from those of lung nodules and the surrounding background breast parenchyma. Therefore, it is not clear how CAD cuing may affect the performance of radiologists in mammography.

The purpose of our study was to assess the performance of radiologists in the detection of masses and microcalcification clusters on digitized mammograms in a CAD environment after modulating cuing sensitivity levels and false-positive rates.

MATERIALS AND METHODS

Seven board-certified radiologists (including M.A.G., C.A.B., C.M.H., L.A.H., T.S.C.) with a minimum of 3 years experience in the interpretation of mammograms participated in this observer performance study. None of the seven observers had participated in the case selection process. All images used in this study were selected from a large and diverse image database established at Magee Womens Hospital, with institutional review board approval and exemption of patient consent. The original database contained mammograms that were collected mainly from several thousand patients undergoing routine mammographic screening at three medical centers (27).

All positive masses were verified at biopsy. All negative cases were rated by radiologists according to the level of concern by using standard Breast Imaging Reporting and Data System, or BI-RADS, recommendations. The negative cases had been diagnosed during at least two subsequent follow-up examinations. Although we routinely acquire four images in a single examination (two views of each breast), for some cases in our digitized database, we have only two images of one breast due to a variety of clinical reasons. By using an established digitization protocol, all mammograms were digitized with a laser-film digitizer (Lumisys, Sunnyvale, Calif), with a pixel size of 100×100 μm and 12-bit digital-value resolution. The quality of the digitizer was monitored routinely to ensure that in the optical density range of 0.2–3.2, digital values were linearly proportional to optical densities (28).

The selection of subtle or difficult cases included several steps. First, we selected a large set of positive cases (200 in this experiment) for which the output scores generated by the CAD scheme were low for the likelihood that the abnormality in question was present (27). Similarly, we used a set of suspicious negative cases (80 in this experiment) for which CAD scores were high for the likelihood that a mass or a cluster of microcalcifications or both were present. Then, two experienced observers pruned the data set by means of visual inspection on the same display as that used in the study with the "true diagnosis" to select the final 120 cases. The total number of positive cases was selected to include a reasonable mix of benign and malignant cases of single and multiple abnormalities, with a minimum of 25 malignant cases of each of the abnormalities.

The resources that were required, in terms of radiologist effort (reading time), were a factor in limiting the number of cases to 120 and the reading modes to five. In 85 cases, mammograms depicted either masses or clusters of microcalcifications or both, and 35 cases were negative for these abnormalities. In 10 of the positive cases, both a mass and a microcalcification cluster were depicted. In all other positive cases, only one abnormality (either a mass or a cluster) was depicted. Hence, the positive cases consisted of 38 verified microcalcification clusters and 57 verified masses. Biopsy results indicated that 27 of clusters and 39 of masses were malignant, while the remaining 11 clusters and 18 masses were benign. Since we were interested in the detection (not classification) of abnormalities, cases were selected on the basis of subtleness of the depicted abnormality, and no attempt was made to balance the number of benign and malignant cases in the dataset. Although study findings suggested that to preserve subtle microcalcifications, mammograms should be digitized with pixel sizes of 50×50 μm or less (15,29), all microcalcification clusters in this study were detectable with our CAD scheme. In addition, we verified that all clusters were visible on images that were digitized with 100×100 μm pixel size.

In this study, radiologists were asked to detect masses and microcalcification clusters on digitized mammograms displayed on a monitor. In most of the 120 cases ($n = 89$), two contralateral images (the same view of left and right breasts) were displayed on the monitor side by side. For some cases ($n = 31$), only a single image was displayed. The latter group was selected from the cases in our database for which we have only two views of one breast. Hence, only one view was displayed in this study, following our study protocol. Table 1 summarizes by type and verified finding the distribution of the abnormalities depicted in the 120 cases. The observers interpreted each case only on the basis of the images displayed on the monitor. No images from previous examinations or other clinical information about the patients was made available during the interpretation.

Each radiologist interpreted the same 120 cases five times by using five display modes. Suspicious regions, as identified with our CAD schemes, were cued on the images in all modes, with the exception of the first mode, in which no CAD results were provided to the radiologists. Two true-positive cuing sensitivity levels (90% and 50%) and two false-positive cuing rates (0.5 or 2.0 per image) were used in these four cuing modes (Table 2). During the cuing modes, when a new case was loaded into the display, radiologists viewed the cued images first. Then they could remove the prompts from the display or add them back at their discretion.

To generate the cues, CAD schemes developed by our group (27) were applied to these 209 images (or 120 cases). The

TABLE 1
Number of Mammographic Cases in Different Categories

Cases	No. of Masses		No. of Microcalcification Clusters		No. of Masses and Clusters		No. of Negative Cases	Total Cases
	M	B	M	B	M	B		
Single-image	10	1	11	3	1	1	4	31
Two-image	20	16	7	7	8	0	31	89
Total	30	17	18	10	9	1	35	120

Note.—B = benign, M = malignant.

TABLE 2
CAD Cuing Conditions of the Five Display Modes

Reading Mode	CAD Cuing	Cuing Sensitivity	Cuing False-Positive Rate
1	No	Not applicable	Not applicable
2	Yes	0.9	0.5
3	Yes	0.9	2.0
4	Yes	0.5	0.5
5	Yes	0.5	2.0

schemes use filtering, subtraction, and topographic region growth algorithms to identify suspicious regions, including masses and microcalcification clusters (30,31). Then, by using nonlinear multilayer multifeature analyses, two artificial neural networks, which have been optimized in our previous studies and reported before (32), were used to classify each region as positive or negative for the presence of an abnormality in question. One network was designed to assess regions suspicious for masses, and the other was for microcalcification clusters. Before applying the artificial neural networks, the schemes initially identified 133 suspicious regions for microcalcification clusters and 831 for masses. Of the 133 clusters, 38 represented true clusters and 95 were false identifications (or a rate of 0.45 [95 of 209 mammograms] false-positive detections per image). Of the 831 mass regions, 57 were true-positive and 774 were false-positive (or 3.7 per image, or 774 of 209 mammograms). The artificial neural networks were then applied to classify all of these regions. Each suspicious region received a likelihood score (from 0 to 1) for being positive. The larger the score, the more likely the region was to represent a true-positive region.

Selection of true-positive and false-positive cues for each display mode was performed separately. Two cuing sensitivities (90% and 50%) were applied to masses and microcalcification clusters. Each abnormality was assigned a number (eg, 1–57 for masses or 1–38 for clusters). A computer program randomly selected the regions to be cued until the required number was reached for the sensitivity level being evaluated. In display modes 2 and 3, with the cuing sensitivity set at 90%, 51 of 57 true masses and 34 of 38 clusters were selected. In modes 4 and 5, with the cuing sensitivity set at 50%, 29 of 57 masses and 19 of 38 clusters were selected. Two false-positive cuing rates (approximately 0.5 and 2.0 false-positive regions per image) were used. Because the number of false-positive clusters identified with the scheme was 95, all of these regions were used in display modes 3 and 5, which provided a false-positive cuing rate of 0.45 (95 of 209 mammograms). In modes 2 and 4, the total false-positive desired cuing rate was 0.5 per image, which was one-fourth of that in modes 3 and 5. Hence, one-fourth of the available false-positive clusters (24 of 95) were selected on the basis of artificial neural network-generated scores, with the 24 highest scoring regions being selected in descending order and resulting in a cuing rate of 0.11 (24 in 209 mammograms).

To reach the overall target of 0.5 and 2.0 false-positive cuing rates per image (including both mass and microcalcification cluster regions), 774 false-positive mass regions were also sorted on the basis of the artificial neural network–generated scores. Then, 82 of the highest scoring false-positive regions were selected from the list for display in modes 2 and 4, and 324 false-positive masses were selected for display in modes 3 and 5. Thus, the false-positive cuing rates for mass only were 0.39 (82 in 209 mammograms) and 1.55 (324 in 209 mammograms) per image, respectively. In summary, modes 2 and 4 included 106 false-positive cues (or 0.5 per image), and modes 3 and 5 included 419 false-positive cues (or two per image).

Each of the 20 reading sessions for individual observers included 30 randomly selected cases that used one reading mode. To eliminate the potential for learning effects, the order of display modes (or cuing rates) for each observer was preselected by using a counterbalanced approach. The 20 sessions were divided into four blocks, with five sessions each. In each block, one observer read five sessions with five different modes in random. However, at each session number in the series (eg, session 6), at least five observers read with different modes, and no more than two readers read with the same mode. For example, in the first session for all the observers, observers started reading with different modes. Because there were seven observers and five display modes, observers 1–5 read with modes 1–5, respectively, while observer 6 read with mode 3 and observer 7 read with mode 2. Last, a study management program was used to randomly select the cases and their sequential order in each session. The random "seed" used in the program was date dependent. Because each observer had a different reading schedule, the cases selected in each session (eg, session 4) and their sequential order for each observer were different. A minimum time delay (10 days) between the two consecutive readings of the same case was implemented.

A standard landscape workstation (Sparc 20; Sun Microsystems, Mountain View, Calif) was used to display the images. Images were not preprocessed, but we did optimize the contrast of each image by means of window and level manipulation for optimal visual display. The image parameters were then fixed. The observers could not manipulate the contrast and brightness settings during the readings. Initially, images were displayed on the screen as subsampled (ie, at low spatial resolution) to fit the screen (with approximately 1,200 × 850 pixels). With zoom and roam functions, the radiologists were able to view the images at full spatial resolution by clicking the appropriate control button or scroll bars. A "Display/Remove" button could be used to superimpose or delete the CAD

cues on the images. Radiologists could make diagnostic decisions while viewing either subsampled or full–spatial-resolution images.

Observers were asked to perform and score two separate tasks. First, they were asked to identify (detect) suspicious areas for the presence of an abnormality and then classify the suspected abnormality as benign or malignant. Once a radiologist pointed to and clicked the cursor on the center of a suspected abnormality, a scoring window appeared, followed by a confidence-level sliding scale. The program automatically recorded all of the diagnostic information entered by the radiologist, including the type of detected abnormality (mass or microcalcification cluster), location (the center of the detected region), and two estimated likelihood scores (from 0 to 1) for the detection (presence or absence) and classification (benign or malignant) of any identified region that was suspected of an abnormality. The likelihood scores were used to generate the free-response receiver operating characteristic curves.

The results of each observer, abnormality, and display mode were qualitatively viewed, and free-response receiver operating characteristic curves were plotted for individual readers and modes, as well as for pooled confidence ratings for all readers since their general patterns were consistent. For testing the hypothesis of equality of the free-response receiver operating characteristic curves (or the detection sensitivities at the same false-positive rates) across four CAD cuing modes, we compared sensitivities among the curves at 10 false-positive rates that were uniformly distributed over the measured range. Sensitivity levels across modalities were compared by using a repeated measures logistic regression model, where the binary outcome variable was replicated over patients, and the independent variables included reader and modality. Estimation was done by using a Generalized Estimating Equation approach (33).

In addition, we analyzed the changes in performance indices (ie, the number of missed true-positive regions in the cued or noncued areas) for the two sensitivity levels (50% and 90%) and the two false-positive cuing rates (0.5 and 2.0 per image). The hypotheses of the equality of the number of missed abnormalities were also tested by using a repeated measures logistic regression, with reader and modality in the model. To examine potential biases for reading the same case five times, the reading results were reordered and analyzed for all cases that were read the first time (regardless of mode) as one group and the second time as another groups, and so on. Performance curves were computed separately for these five mutually exclusive groups and were compared by using the analysis of variance test.

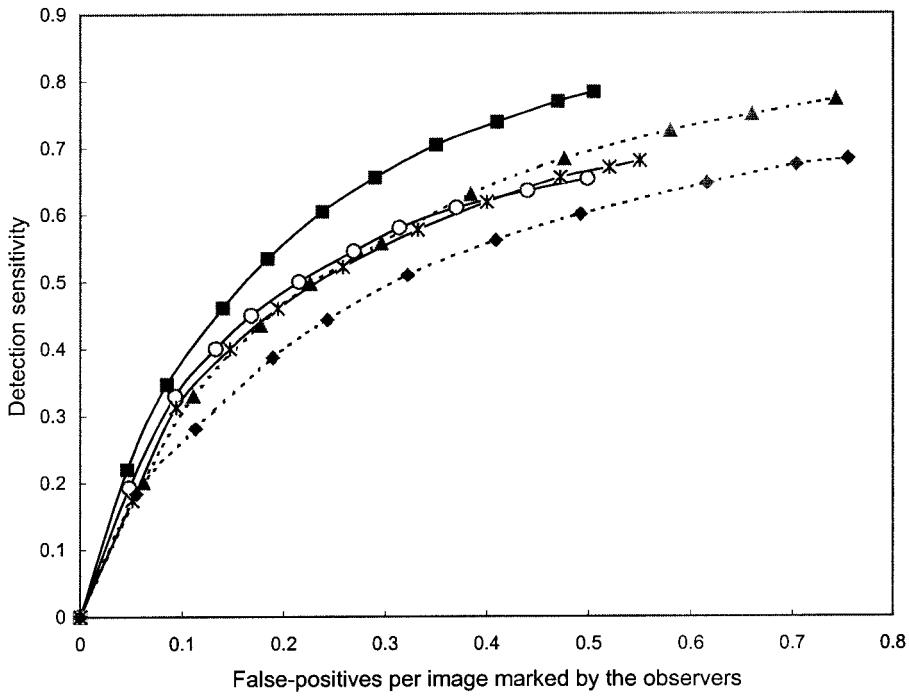

Figure 1. Free-response receiver operating characteristic curves for the average detection of mammographic abnormalities (including both masses and microcalcification clusters) by seven participating radiologists using five display modes. ○ = mode 1, ■ = mode 2, ▲ = mode 3, ∗ = mode 4, and ◆ = mode 5.

RESULTS

Performance curves varied among observers, but the general pattern was consistent. Figures 1–3 demonstrate curves of the average performance of the seven observers for the detection of either abnormality, masses, or microcalcification clusters, respectively. As can be noted from the noncued results (mode 1), the task in general was challenging because of the display environment, the subtlety of the abnormalities, or both.

Figure 1 demonstrates that both sensitivity and specificity of the CAD results affected observer performance. The differences among modes 2–5 were highly significant ($P < .01$). However, the results showed different patterns for the detection of masses compared with microcalcifications. In the case of masses (Fig 2), specificity of the CAD results (or cuing false-positive rate) affected the observers in a more significant manner. The differences among modalities were statistically significant ($P < .01$), with the performance decreasing as the number of cued regions increased. In the case of clusters (Fig 3), observer performance was affected to a greater extent by the cuing sensitivity. The combination of case subtlety and viewing of soft copies rendered the test of microcalcification cluster detection so difficult that only approximately 60% were detected without cuing or with cuing at low sensitivity (modes 4 and 5). With the support of highly sensitive cues, the performance improved to a detection rate of approximately 75% ($P < .01$).

Highly accurate cuing (ie, 90% sensitivity and 0.5 false-positive cue per image) helped the observers to improve their performance, compared with the noncued environment ($P < .01$). As the accuracy of the cuing decreased, so did the performance of the typical observer. This effect continued for either detection task, but the detection of microcalcification clusters was more significantly affected by sensitivity of the cuing in our case. Most important, perhaps, our study results clearly indicate that poorly performing CAD (Fig 1) can result in significant degradation of observer performance ($P < .01$).

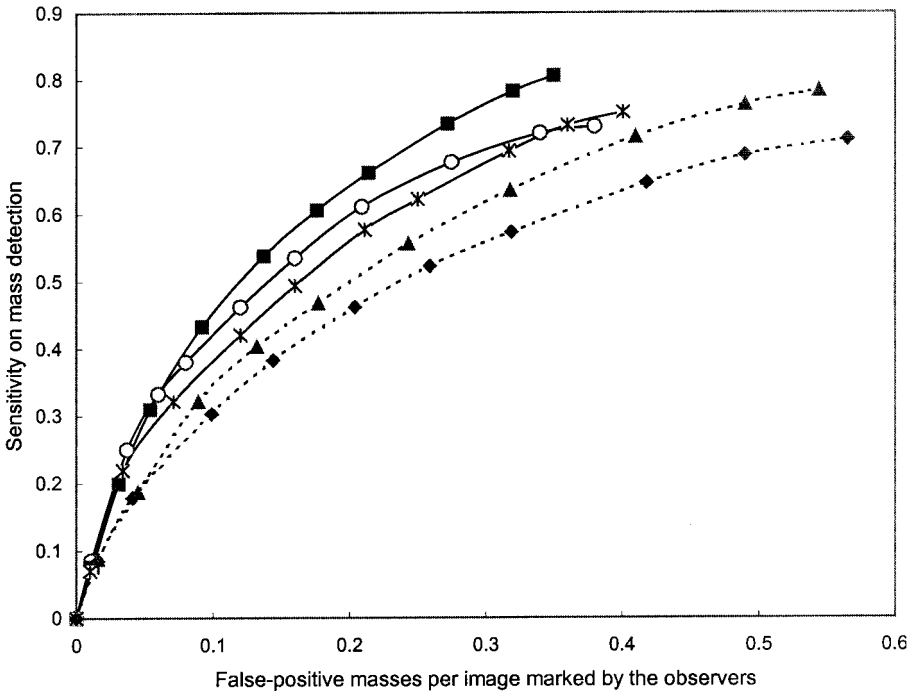

Figure 2. Free-response receiver operating characteristic curves for the average mass detection by seven radiologists using five display modes. ○ = mode 1, ■ = mode 2, ▲ = mode 3, ∗ = mode 4, and ◆ = mode 5.

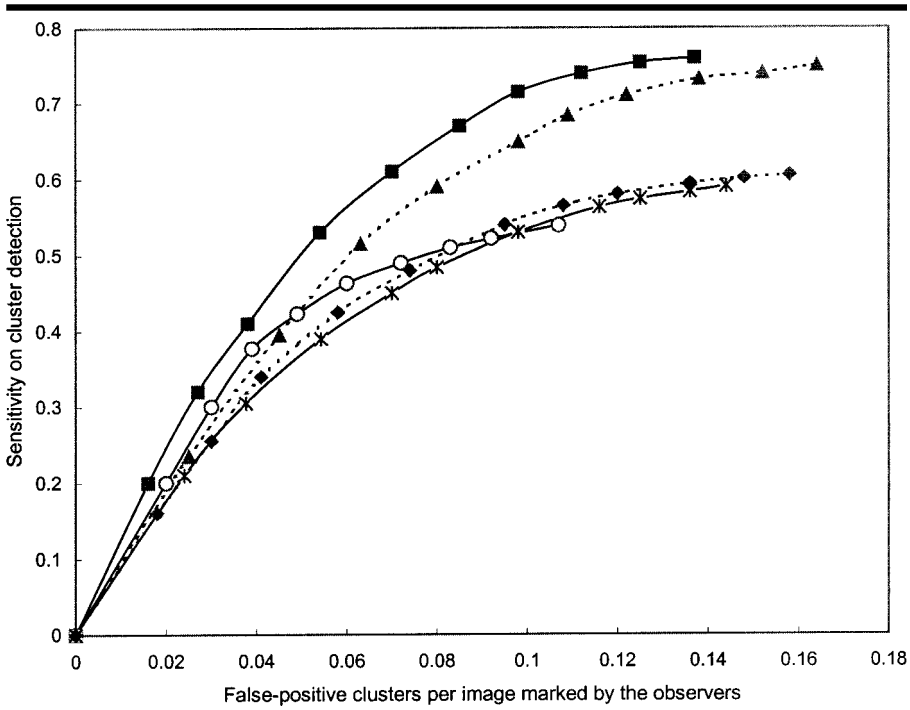

Figure 3. Free-response receiver operating characteristic curves for the average microcalcification cluster detection by seven radiologists using five display modes. ○ = mode 1, ■ = mode 2, ▲ = mode 3, ∗ = mode 4, and ◆ = mode 5.

Table 3 demonstrates the number of CAD-cued abnormalities that were identified by each radiologist in mode 1 (noncuing) but were missed in other (cued) modes. Some increases in rejection rates of true-positive regions were observed when the number of cues increased, but the results were not significant ($P > .05$).

Table 4 summarizes the number of missed abnormalities in noncued areas during CAD-cued observations. The table data show that for the highly sensitive cuing modes (eg, modes 2 and 3, where only 10% of true-positive regions were not cued), the majority of missed abnormalities (>94%) were also missed in mode 1. As CAD cuing sensitivity was reduced to 50%, the average number of missed abnormalities in noncued areas increased significantly ($P < .05$). More important, approximately 30% of these regions were detected by the radiologists in mode 1. The increase of the false-positive cuing rate from 0.5 to 2.0 per image (mode 4 vs mode 5, respectively) increased the number of missed abnormalities in noncued areas, from an average of 14.4 to 18.0, which was not significant ($P = .16$) and most likely due to the small sample size. In this case, the observers also missed significantly more regions that were detected in mode 1 ($P = .03$). In general, the number of missed abnormalities (false-negative rate) in the noncued areas increases as the cuing sensitivity decreases and the false-positive cuing rate increases. As a result, mode 5 had the highest miss rate in noncued areas. When we compared detection performances for benign and malignant abnormalities, the latter group was somewhat better detected (probably due to differences in subtleness), but the differences between modes were similar to those of the benign group.

The pooled classification confidence ratings (malignant vs benign) provided by the seven observers on all identified true-positive regions for each mode were used to generate and compare the area under the receiver operating characteristic curve (A_z) values for the different modes (ROCFIT; Metz CE, Herman BA, Shen JH, University of Chicago, Il) (34). A_z values were estimated by using maximum likelihood estimation under the binormal assumption. The A_z values for the classification performance over all readers were 0.70 ± 0.02, 0.69 ± 0.02, 0.69 ± 0.02, 0.70 ± 0.02, and 0.68 ± 0.02 for modes 1–5, respectively. Comparison of each pair of modes did not result in any significant differences ($P > .05$). Hence, once the abnormality was identified (detected), the ability of the observer to distinguish between benign versus malignant abnormalities (classification) was not significantly affected ($P > .05$) by the cuing mode or lack thereof. Although there were differences in performance

among the observers, we did not identify any correlation of either the detection or classification tasks with observer experience, as measured by the number of years of interpreting mammograms or the average number of mammograms interpreted per year. The performance trends we observed were consistent for all observers.

The minimum time delay between two consecutive readings of the same case by the same observer was set at 10 days, but the actual time delay ranged from 12 to 154 days, with an average time delay of 48 days. When we examined the results after reordering the cases by their order of appearance (ie, first time, second time, etc), regardless of the mode, no significant ($P > .8$) difference between the groups was identified (Fig 4). Similar performance patterns were observed when 31 cases that included only one image were excluded from the analyses, and the detection results were not significantly altered in any comparison between those for the whole group (120 cases) and the subset of 89 cases containing two images ($P > .5$).

TABLE 3
Number of Missed Abnormalities Identified as Suspicious in Mode 1 (Noncued) but Missed in Other Modes Despite the Fact that the Abnormality in Question Was Cued

Reader	Mode 2	Mode 3	Mode 4	Mode 5
1	5	5	3	3
2	5	4	4	3
3	5	6	3	6
4	3	1	5	4
5	1	9	5	11
6	5	4	8	5
7	3	1	4	2
Average	3.9	4.3	4.6	4.9

TABLE 4
Number of Missed Abnormalities in Noncued Regions

Reader	Mode 2	Mode 3	Mode 4	Mode 5
1	5 (1)	5 (1)	13 (3)	14 (5)
2	6 (0)	8 (0)	19 (2)	21 (7)
3	5 (1)	5 (0)	11 (2)	15 (3)
4	5 (0)	6 (0)	19 (3)	25 (5)
5	6 (0)	4 (0)	10 (4)	13 (5)
6	7 (1)	7 (2)	14 (4)	20 (9)
7	6 (0)	5 (0)	15 (3)	18 (6)
Average	5.7 (0.4)	5.7 (0.4)	14.4 (3.0)	18.0 (5.7)

Note.—Data in parentheses are the number of missed regions that were detected in mode 1 (noncued).

DISCUSSION

This preliminary study has to be clearly viewed as a study performed under laboratory conditions. Before any generalization of the results is contemplated, it has to be considered that conditions in this study were removed from the typical clinical environment. However, the consistency of the patterns observed for the individual readers and the group as a whole warrant further assessment of the affect of CAD performance on the observer.

Clearly, the expectation that observers can readily and easily discard most false-positive cues regardless of their presentation or prevalence was not what we found (14). Both true- and false-positive cues affected the results. The effect was also dependent on the type of abnormality and its subtleness (detection difficulty). Despite significant reader, case, and mode variability, the results we obtained were consistent and interpretable. As expected, at low specificity levels, all CAD-cued modes aid in increasing sensitivity of observers, as can be seen from the tendency to cross the noncuing performance curve. This observation is consistent with some of the results previously reported by others, but it may not be clinically relevant in situations in which most abnormalities are not as difficult to detect as those in this study.

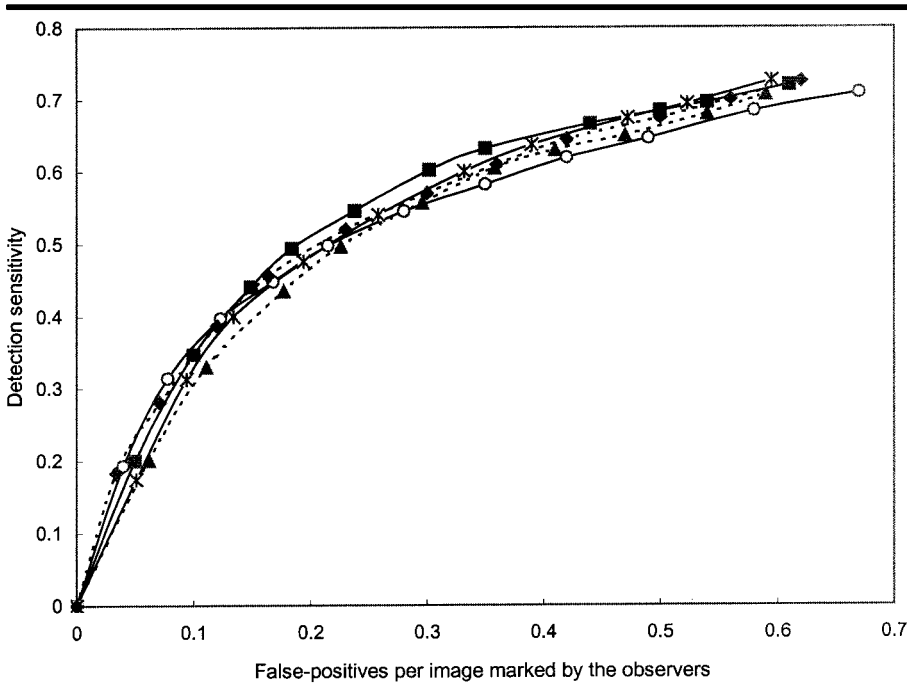

Figure 4. Free-response receiver operating characteristic curves for the average detection of abnormalities by seven radiologists as a function of the order of appearance: ○ = first time, ■ = second time, ▲ = third time, ∗ = fourth time, and ◆ = fifth time, regardless of the reading mode.

Our results suggest that the use of a CAD-cued environment during the interpretation of mammograms has to be carefully investigated and fully understood before it is widely accepted in a routine clinical practice. In particular,

one should consider the cuing performance level of the scheme itself and the potential increase in missed abnormalities in noncued regions, because the possible liability associated with false-negative interpretations far exceeds that of false-positive readings (26).

The general consistency of our results is somewhat surprising in view of the fact that cuing rates were maintained only for short durations (within a single session of 30 cases). Unlike the display environment, the CAD results in our study emulated what can be expected by using current levels of CAD performances, as well as what one hopes to achieve by using CAD in the future. The range of CAD performances that were used for cuing at 90% sensitivity at 0.5 false-positive identification per image to 50% sensitivity at two false-positive identifications per image clearly makes this study interesting in enabling an assessment of what could be expected with improved CAD results. It is interesting to note that for all display modes, the use of CAD cuing with either high or low performance had a limited effect on observers when they operated at a conservative level. Namely, they indicated only regions they were confident about, and, therefore, they had low false-positive rates. This stemmed largely from the fact that the CAD cuing depicted mainly areas on the image that were truly appropriate (reasonable) as suspicious. As observers loosened their criteria (ie, indicated a larger number of suspicious regions), the CAD-cuing performance affected observers in a more significant manner. Namely, the use of a better performing cuing scheme significantly improved observer performance, while the use of poorly performing cuing schemes significantly degraded observer performance.

Analysis of the data sets after the reorder of cases by appearance indicates that learning effects, if any, were not a significant factor in this study. Although all selected abnormalities in this study were detectable with CAD schemes and visible on displayed images, the relatively low detection levels of the seven participating observers in the case of subtle clustered microcalcifications suggest that this task is likely to be a continuing challenge when soft copy is used for this purpose. We are not aware of any comprehensive study in which this issue was assessed, and our results, albeit preliminary, suggest that such a study should be performed.

Despite the limited information (no prior studies or reports and only a single view for each breast) and the fact that different abnormalities were detected in each mode, the classification performances of determining that an identified abnormality was either benign or malignant were reasonable and consistent. It was encouraging to learn that once detected, the task of classifying the abnormality as benign or malignant was not affected by the detection cuing performance, which points to the fact that these are likely to be two distinct and largely independent tasks. Our CAD scheme was designed solely for detection purposes. Other classification schemes (12) have been shown to perform well, and, when used during interpretation, significantly improved tissue classification performance of the observers (10,11).

The overall detection sensitivity of the radiologists was in general relatively low compared with that observed in the clinical environment. This may be due to the fact that most of the cases selected for this study were subtle, and reading was performed on soft copy by using a limited number of views without prior examinations being available for comparison. We note a difference between this and other reported studies (14,15) where observers could view both film hard-copy images and low-spatial-resolution soft-copy images with CAD-cued areas on the screen. Not providing film hard-copy images to the observers could have been a significant factor in lowering detection sensitivity in this study. This resulted in a crossing of the performance curves for the detection of microcalcifications (Fig 3), since the noncued mode exhibited a "capping" effect (an imposed upper limit) that was removed with the aid of CAD cuing. This does not invalidate any of the analyses or observations made in this study. Despite the generally low level of performance and the high prevalence of abnormalities in our data set, we believe that on a relative scale, the results concerning the general trends we observed are valid. We emphasize that our study design called for a change in mode (hence, abnormality rates) at each session. The effects we observed under these conditions are probably different and likely minimized, as compared with those in a study design in which each mode is read to its completion before any prevalent changes (ie, change to a different mode).

In conclusion, our preliminary study results indicate that in a laboratory environment, observer performance in the detection of subtle mammographic abnormalities is significantly affected by the inherent performance of a cuing system. High-performance cuing systems can significantly improve observer performance. On the other hand, low-performance cuing systems can significantly degrade observer performance. These findings, together with the intermode consistency we observed, are important, since there could be diagnostic implications associated with the inappropriate use of or reliance on CAD results during the interpretation. These issues have to be further investigated with larger data sets and a more closely simulated clinical environment.

References

1. Mettlin C. Global breast cancer mortality statistics. CA Cancer J Clin 1999; 49:135–137.
2. Smith RA. Breast cancer screening among women younger than age 50: a current assessment of the issues. CA Cancer J Clin 2000; 50:312–336.
3. Feig SA, D'Orsi CJ, Hendrick RE. American College of Radiology guidelines for breast cancer screening. AJR Am J Roentgenol 1998; 171:29–33.
4. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology 1992; 184:613–617.
5. Thurfjell EL, Lernevall KA, Taube AS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241–244.
6. Vyborny CJ, Giger ML. Computer vision and artificial intelligence in mammography. AJR Am J Roentgenol 1994; 162: 699–708.
7. Hoffman KR. In the next decade automated computer analysis will be an accepted sole method to separate "normal" from "abnormal" radiological images. Med Phys 1999; 26:1–4.
8. Nishikawa RM, Giger ML, Schmidt RA, Wolverton DE, Collins SA, Doi K. Computer-aided diagnosis in screening mammography: detection of missed cancers (abstr). Radiology 1998; 209(P):353.
9. Nawano S, Murakami K, Moriyama N, Kobatake H. Computer-aided diagnosis in full digital mammography. Invest Radiol 1999; 34:310–316.
10. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 1999; 6:22–33.
11. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 1999; 212:817–827.
12. Leichter I, Fields S, Nirel R, et al. Improved mammographic interpretation of masses using computer-aided diagnosis. Eur Radiol 2000; 10:377–383.
13. Thurfjell E, Thurfjell MG, Egge E, Bjurstam N. Sensitivity and specificity of computer-assisted breast cancer detection in mammography screening. Acta Radiol 1998; 39:384–388.
14. Doi T, Hasegawa A, Hunt B, Marshall J, Rao F, Roehrig J. Clinical results with the R2 ImageCheck Mammographic CAD

system. In: Doi K, MacMahon H, Giger ML, Hoffman KR, eds. Computer-aided diagnosis. Amsterdam, the Netherlands: Elsevier Science, 1999; 201–207.
15. Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000; 215: 554–562.
16. Sittek H, Perlet C, Helmberger R, Linsmeier E, Kessler M, Reiser M. Computer-assisted analysis of mammograms in routine clinical diagnosis. Radiologe 1998; 38:848–852. [German]
17. Funovics M, Schamp S, Lackner B, Wunderbaldinger P, Lechner G, Wolf G. Computer-assisted diagnosis in mammography: the R2 ImageCheck System in detection of speculated lesions. Wien Med Wochenschr 1998; 148:321–324. [German]
18. Nishikawa RM, Yarusso LM. Variations in measured performance of CAD schemes due to database composition and scoring protocol. Proc SPIE Medical Imaging Conference 1998; 3338:840–844.
19. Brake GM, Karssemeijer N, Hendriks JH. Automated detection of breast carcinomas not detected in a screening program. Radiology 1998; 207:465–471.
20. Gray JE. Against the proposition, at point/counterpoint of in the next decade automated computer analysis will be an accepted sole method to separate "normal" from "abnormal" radiological images. Med Phys 1999; 26:3–4.
21. King M, Stanley GV, Burrows GD. Visual search in camouflage detection. Hum Factors 1984; 26:223–234.
22. Krose BA, Julesz B. The control and speed of shifts of attention. Vision Res 1989; 29:1607–1619.
23. Parker TW, Kelsey CA, Moseley RD, Mettler FA, Garcia JF, Briscoe DE. Directed versus free search for tumors in chest radiographs. Invest Radiol 1982; 17:152–155.
24. Kundel HL, Nodine CF, Krupinski EA. Searching for lung nodules: visual dwell indicates locations of false-positive and false-negative decisions. Invest Radiol 1989; 24:472–478.
25. Nodine CF, Kundel HL, Toto LC, Krupinski EA. Recording and analyzing eye-position data using a microcomputer workstation. Behav Res Methods Instrum Comput 1992; 24:475–485.
26. Krupinski EA, Nodine CF, Kundel HL. Perceptual enhancement of tumor targets in chest x-ray images. Percept Psychophys 1993; 53:519–526.
27. Zheng B, Sumkin JH, Good WF, Maitz GS, Chang YH, Gur D. Applying computer-assisted detection schemes to digitized mammograms after JPEG data compression: an assessment. Acad Radiol 2000; 7:595–602.
28. Zheng B, Chang YH, Gur D. On the reporting of mass contrast in CAD research. Med Phys 1996; 23:2007–2009.
29. Chan HP, Niklason LT, Ikeda DM, Lam KL. Digitization requirements in mammography: effects on computer-aided detection of microcalcifications. Med Phys 1994; 21:1203–1211.
30. Zheng B, Chang YH, Staiger M, Good WF, Gur D. Computer-aided detection of clustered microcalcifications in digitized mammograms. Acad Radiol 1995; 2:655–662.
31. Zheng B, Chang YH, Gur D. Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis. Acad Radiol 1995; 2:959–966.
32. Zheng B, Chang YH, Good WF, Gur D. Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme. Acad Radiol 1997; 4:497–502.
33. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73:13–22.
34. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat Med 1998; 17: 1033–1053.