Improvement in Radiologists?? Detection of Clustered Microcalcifications on Mammograms

Article *in* Investigative Radiology · November 1990

DOI: 10.1097/00004424-199010000-00006 · Source: PubMed

9 authors, including:

Heang-Ping Chan
University of Michigan
541 PUBLICATIONS **11,734** CITATIONS

SEE PROFILE

Heber Macmahon
University of Chicago
369 PUBLICATIONS **14,027** CITATIONS

SEE PROFILE

Improvement in Radiologists' Detection of Clustered Microcalcifications on Mammograms

The Potential of Computer-Aided Diagnosis

HEANG-PING CHAN, PHD, KUNIO DOI, PHD, CARL J. VYBORNY, MD, PHD, ROBERT A. SCHMIDT, MD, CHARLES E. METZ, PHD, KWOK LEUNG LAM, PHD, TOSHIHIRO OGURA, BS, YUZHENG WU, BS, AND HEBER MACMAHON, MD

Chan H-P, Doi K, Vyborny CJ, Schmidt RA, Metz CE, Lam KL, Ogura T, Wu Y, MacMahon H. Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis. Invest Radiol 1990;25:1102–1110.

Relatively simple, but important, detection tasks in radiology are nearing accessibility to computer-aided diagnostic (CAD) methods. The authors have studied one such task, the detection of clustered microcalcifications on mammograms, to determine whether CAD can improve radiologists' performance under controlled but generally realistic circumstances. The results of their receiver operating characteristic (ROC) study show that CAD, as implemented by their computer code in its present state of development, does significantly improve radiologists' accuracy in detecting clustered microcalcifications under conditions that simulate the rapid interpretation of screening mammograms. The results suggest also that a reduction in the computer's false-positive rate will further improve radiologists' diagnostic accuracy, although the improvement falls short of statistical significance in this study.

Key words: computer-aided diagnosis; missed diagnoses; breast radiography; receiver operating characteristic (ROC) study.

MANY MISSED RADIOLOGIC diagnoses can be attributed to human factors such as subjective or varying decision criteria, distraction by other image features, or simple oversight.[1–4] Studies suggest that these errors may be inevitable with human observers and that they are not strongly related to experience.[5,6] Although longer viewing times or double readings may improve radiologists' performances, these strategies do not eliminate errors completely.[7–10] In interpreting mammograms, as with other examinations, radiologists do not detect all carcinomas that are visible on retrospective analyses of the images.[1,6,11–15] Although these failed detections are often a result of the very subtle nature of the radiographic findings, oversight by the radiologist also contributes to missed diagnoses.[1,11]

We have developed a computer program that can automatically locate clustered microcalcifications on mammograms.[16,17] The program detects microcalcifications quite sensitively and is intended to aid radiologists by identifying locations of suspicious clusters, rather than by estimating the probability of malignancy for any given cluster. As such, the program can potentially decrease the miss rate of radiologists, particularly under circumstances in which a large volume of cases are reviewed, such as in the interpretation of screening mammograms. We conducted a receiver operating characteristic (ROC) study that compared human detection performance with and without the aid of the computer under conditions that simulated this environment. We also analyzed the effect of the computer's false-positive detection rate on the radiologists' accuracy in identifying true clusters.

Materials and Methods

Selection of the Case Sample

Sixty screen-film mammograms were selected for this study by an experienced mammographer, who did not participate in the observer performance experiments. A single craniocaudal or mediolateral oblique view was chosen from each patient file. Thirty of the mammograms contained a single cluster of subtle microcalcifications that had been verified by biopsy or by magnification views. These films were selected from a much larger file contain-

Presented at the 74th Scientific Assembly and Annual Meeting of the Radiological Society of North America, November 1988, Chicago, Illinois.

From the Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, The University of Chicago, Illinois.

Reprint requests: Heang-Ping Chan, PhD, Department of Radiology, The University of Michigan, Taubman Center 2910A—Box 0326, Ann Arbor, MI 48109-0326.

Received December 4, 1989, and accepted for publication May 1, 1990.

ing difficult clustered microcalcifications accumulated over the last five years at the University of Chicago. A cluster was chosen only if it was not readily visible on casual inspection of the film. Each cluster contained four or more microcalcifications, and the majority of the individual microcalcifications had a diameter less than 0.3 mm. The subtlety of the cases chosen is demonstrated by Figures 4 and 5, which show two typical cases used in the observer performance study, and also by the estimated contrast of the digitized microcalcifications described in the next section.

The other half of the mammograms used in the study were "normal" cases that were free of clustered microcalcifications on careful inspection. Some of these mammograms showed other abnormal findings, such as areas of increased density or parenchymal distortion, but were classified as "normal" for the purpose of the study if microcalcifications were absent. The 60 mammograms included a typical range of breast sizes and densities. Films in the normal group were chosen to generally match the abnormal group in terms of projection and radiographic technique.

Digitization and Display of Mammograms

High-quality printed digital mammograms were used throughout the observer study. This allowed standardization of the display format when the computer-detected locations of clustered microcalcifications were displayed. All mammograms were digitized with a precision drum scanner system,[18] using a sampling aperture of 0.1 mm × 0.1 mm and a 0.1 mm × 0.1 mm sampling array. The drum scanner was calibrated such that the optical density range from 0.2 to 2.75 was digitized to 1024 gray levels (pixel values). The calibration curve was linear in the density range from 0.4 to 2.2, with a slope of 0.003 optical density unit/pixel value; the slope of the calibration curve decreased gradually outside this density range. With this calibration, the distribution of the contrasts of the individual microcalcifications (defined as the difference between the maximum pixel value of a microcalcification and the average pixel value of its local background) in the 30 abnormal mammograms had a mean of 46 pixel values and a standard deviation of 26 pixel values, corresponding to a contrast of 0.14 ± 0.08 optical density units. The root-mean-square noise in the local background was approximately 13 pixel values. Note that the presence of the random noise generally caused an overestimation of the visual contrasts of the microcalcifications because the maximum pixel value was used in the above definition.

Digital image processing and automated computer detection were performed on a DEC Vax 11/750 computer (Digital Equipment Corp., USA). The digitized mammograms and the computer-detected locations were printed on Konica single-emulsion LP film by a Konica laser printer (Konica, Japan) for the observer performance study. The laser printer had a beam spot size of 0.075 mm × 0.105 mm and a pixel pitch of 0.0875 mm in both directions. Therefore, the viewed digital images were reduced in size by about 12.5% in each dimension when compared with the original mammogram. The large-area contrast of the reconstituted image on film was maintained essentially identical to that of the original mammogram.

Methods for Automated Detection of Clustered Microcalcifications

Our general approach to the automated detection of clustered microcalcifications on mammograms has been described elsewhere.[16,17] Briefly, a digital mammogram is processed by linear or nonlinear methods to improve the signal-to-noise ratio (SNR) of microcalcifications on the image. Gray-level thresholding techniques, which combine a global gray-level thresholding procedure and a locally adaptive gray-level thresholding procedure, are then employed to extract potential signal sites from the noise background. Subsequently, signal-extraction criteria are imposed on the potential signals to distinguish true signals from noise or artifacts. The computer then indicates locations that may contain clusters of microcalcifications on the image.

In this study, we employed a difference-image technique with linear spatial filters for SNR enhancement. The difference filter was composed of a matched filter with a 3 × 3-pixel kernel and a box-rim filter with an outer width of 9 pixels and an inner width of 5 pixels. For local gray-level thresholding, the threshold levels were varied as the number of standard deviations (SD) of the pixel-value variation in a 51 × 51-pixel region around each potential signal site in the filtered image. A potential signal extracted by thresholding was retained as a detected signal if its area was greater than 3 pixels. An area upper bound of 80 pixels and a contrast upper bound of 10 times the SD of the local pixel-value variation were chosen to exclude large-area or high-contrast signals that might be caused by artifacts or by large, benign calcifications. Finally, a clustering procedure was performed for locating clusters that included a minimum of three signals within a 1.2-cm-diameter circular area.[17]

To further improve the accuracy of detection, we included in this study a new signal-extraction criterion not previously reported. For the potential microcalcifications that passed the size and contrast criteria, the power spectrum in an n × n-pixel region centered at each signal site was calculated. Each pixel value of the low-frequency background in this region (Fig. 1B) was estimated by averaging the two pixel values which were obtained by curve fitting with polynomials in the horizontal and vertical directions. The background was subsequently subtracted from the original image region (Fig. 1A). After background correction (Fig. 1C), each n × n-pixel region was Fourier-transformed and its power spectrum (Fig. 1D) calculated. Typical power spectra at true signal and false signal sites are compared in Figure 1. A false signal site generally showed a broader power spectrum than that of a true signal site. We therefore characterized the power spectrum by its first moment, defined as the weighted average of radial spatial frequency over the two-dimensional power spectrum.[19]

Figure 2 shows histograms of the first moment of the power spectrum in a 64 × 64-pixel region for the true and false signals detected in a training set of 47 clinical mammograms that were not used in the observer study. The area under each histogram was normalized to unity. It can be seen that a large fraction of the false signals have first-moment values larger than those of the true signals, probably because the power spectra of the false-signal regions are more dominated by random background noise. This separation was smaller when a region size of 32 × 32 pixels or 128 × 128 pixels was used. We therefore chose a 64 × 64-pixel region for calculation of the power spectrum and adopted 3 cycles/mm as an upper bound for the first moment as a criterion with which to distinguish true signals from false signals.

The 60 mammograms selected for the ROC study were processed with the automated detection program described above; this included the various signal-extraction criteria described elsewhere[17] in addition to the power-spectral criterion just outlined. A region of 800 × 1000 pixels that contained the main breast area and the cluster of microcalcifications (for the abnormal cases) was processed for each image. The results for computer detection alone are plotted as a free response operating characteristic (FROC) curve[20] in Figure 3. This curve shows the fraction of true-positive (TP) cluster detected as a function of the number of false-positive (FP) clusters detected per image when the local threshold level varies. The data points correspond to the detection rates achieved at local threshold levels of 3.2 to 4.2 SD. For the set of mammograms used in the observer study, the TP cluster detection accu-

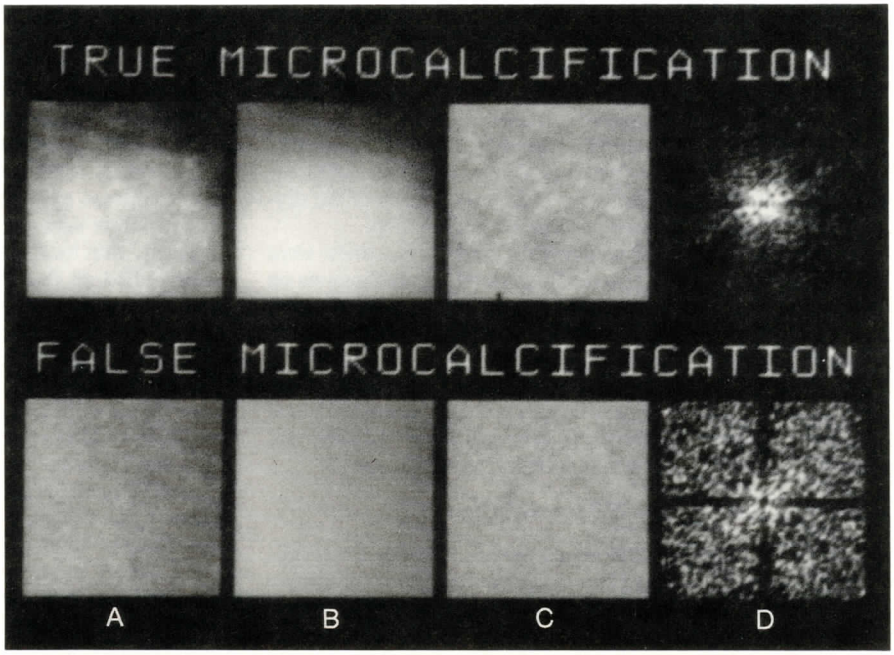

Fig. 1. Determination of the Fourier power spectrum in a region of 64 × 64 pixels centered at a potential signal site for the first-moment test. Upper row: true microcalcification. Lower row: false microcalcification. (A) Unprocessed image region; (B) estimated low-frequency background by curve fitting; (C) background-corrected region; (D) power spectrum of Figure 1C normalized and plotted in log scale.

racy of our automated detection program reaches 87% at an FP detection rate of 4 clusters per image. An ROC study then was performed to determine whether this performance level could result in an improvement in radiologists' performance when the CAD results were displayed on images.

Observer Performance Studies

The ROC study was designed to allow comparison of radiologists' accuracies in the detection of microcalcifications when the digitized mammograms were read in each of the three following conditions: (1) A single mammogram was interpreted without CAD. (2) A pair of identical mammograms was presented side-by-side, with circles superimposed on one film to indicate locations at which the computer program detected clustered microcalcifications. The computer's accuracy in this condition was 87%—TP cluster detection with an average of 4 FP clusters per image; this will be referred to as "level 1 CAD" in the following discussion. (3) A pair of identical mammograms was presented in a way similar to condition (2), except that a simulated level of computer

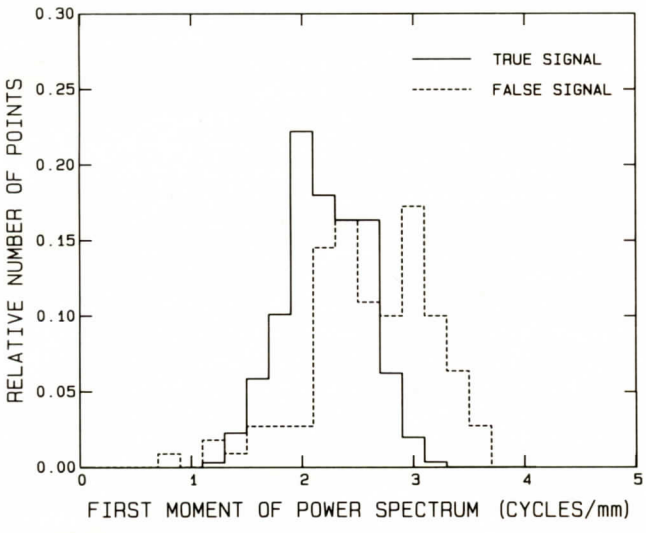

Fig. 2. Histograms of the first moment of the two-dimensional power spectrum in a 64 × 64-pixel region for true and false signals detected in 47 clinical mammograms at a local threshold level of 3.6 SD (standard deviation of the local pixel-value variation).

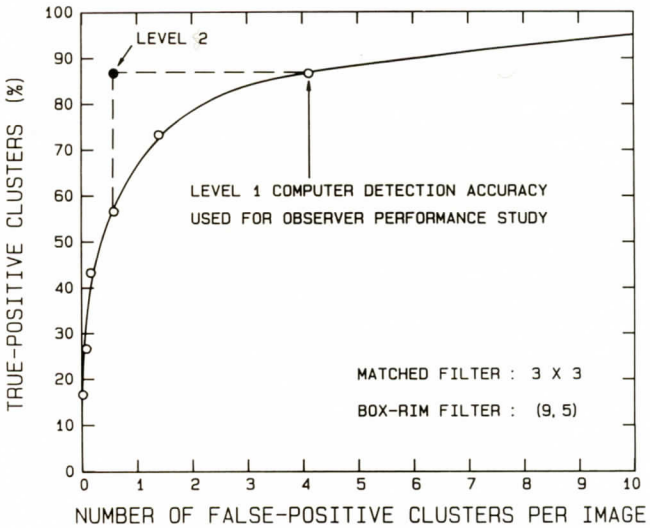

Fig. 3. FROC curve obtained with the computer detection program alone for the 60 clinical mammograms used in the observer performance studies. The data points (open circles) from right to left indicate results obtained with local threshold levels from 3.2 to 4.2 SD in 0.2 SD increments. The result with a true-positive (TP) rate of 87% and false-positive (FP) rate of 4 clusters per image is referred to as "level 1" accuracy in the observer performance study of CAD. A simulated detection level, which combines the TP rate at 3.2 SD and the FP rate at 3.6 SD (solid circle), is referred to as "level 2" accuracy.

accuracy was employed, providing 87% TP cluster detection with an average of 1 FP cluster in every two images; this will be referred to as "level 2 CAD." We included this as-yet-unattained level of computer accuracy to evaluate the effect of the computer's FP detection rate on the human observer's performance.

We prepared three sets of contrast- and density-matched images with the laser printer. The first set of 60 images was printed directly from the digitized image data without computer detection results superimposed. In the second set, circles were superimposed at the computer-detected locations obtained with a local threshold level of 3.2 SD (Fig. 3), ie, with the computer operating at level 1 accuracy. The third set included superimposed circles that corresponded to the level 2 accuracy of computer detection. To produce this simulated level, the computer-detected locations were obtained by using the TP locations from a local threshold level of 3.2 SD and the FP locations from a more stringent local threshold level of 3.6 SD (see also Fig. 3).

Seven attending radiologists and eight radiology residents participated in the ROC studies. For each observer, the ROC study consisted of three reading sessions on separate days. During each session, an observer read one third of the mammograms (20 images) sequentially in each of the three conditions described above. The 20 mammograms to be read in each condition during a session were selected so that half of the cases were normal and the other half were abnormal. Each of the 60 cases was read once in each session. For each observer, the reading order of the 20 mammograms in each group was varied randomly, and the order of the three reading conditions in a session and the grouping of mammograms to be shown in each condition were varied systematically in a way described elsewhere.[21,22] This minimized any potential effects of learning or observer fatigue on the relative ranking of the detection performance achieved with the three reading conditions.

In each reading session, the mammograms were mounted on an alternator, and one case was shown at a time. For each of the two reading conditions with CAD, the films with and without CAD reports were mounted side-by-side, and the observer could choose to read either film first. Reading time was limited to 5 seconds per case for all three reading conditions. A magnifying glass was provided, but its use was optional. The observer was informed of the reading condition to be employed and the level of computer accuracy before each group of 20 cases was presented.

A five-category rating scale was employed to represent each observer's confidence regarding the presence of a cluster of three or more microcalcifications (1 = definitely not present; 2 = probably not present; 3 = possibly present; 4 = probably present; 5 = definitely present). Before the experiment, the observer was informed of the approximate prevalence of abnormal cases in the case samples. They were asked specifically to rate the mammograms in terms of the presence of clustered microcalcifications rather than the likelihood of malignancy. They were asked also to keep their interpretation of the five rating categories constant throughout the three reading sessions of the experiment. A training session was given to each observer before the first reading session to acquaint the observer with the information provided by the CAD method.

To perform ROC analysis with localization (LROC analysis), the observer was asked to circle with a wax pencil the location most suspicious for the presence of a cluster in each case, including the cases rated 1 or 2. For this purpose, each film was overlaid with a piece of clear vinyl sheet. This sheet did not affect the visibility of fine details on the mammograms. After each reading session, the location marked on each film was recorded and the circle was completely erased.

Data Analysis

Receiver operating characteristic analysis[23,24] was employed for comparison of the observers' performance in the detection of clustered microcalcifications on mammograms without and with the two levels of CAD. For the conventional ROC analysis, a binormal ROC curve was fitted to each observer's confidence-rating data from each reading condition by maximum likelihood estimation.[25] The index A_z, which represents the area under the best-fit binormal ROC curve when it is plotted in the unit square, and the slope and intercept parameters of the binormal ROC curve when it is plotted on normal probability scales, were calculated for each fitted curve. The statistical significance of the difference between each pair of reading conditions was analyzed by applying a "two-tailed" Student's t-test for paired data to the observer-specific A_z index values.[22] For each reading condition, the overall performance of the group of observers was summarized by a "composite" ROC curve, which was obtained by averaging the slope and intercept parameters of the individual observer's ROC curves for that reading condition.

Receiver operating characteristic with localization (LROC) curves[26] also were determined for each reading condition. At each confidence-rating level, the vertical coordinate of an LROC curve represents the fraction of images with one local abnormality in which the abnormality was correctly detected and localized by the observer. In our experiment, localization was considered correct if the cluster of microcalcifications was enclosed within a 1-cm to 2-cm-diameter circle drawn by the observer. These true-positive rates with correct localization of cluster and the false-positive fractions (FPFs) of images were calculated at the five confidence-rating levels for each reading condition and each observer. A pooled[23] LROC curve for each reading condition was determined by the true-positive rates and the FPFs at each rating level averaged over the 15 observers.

Results

Figures 4 and 5 show two examples of the abnormal mammograms used in the ROC studies. Figure 4 shows a mediolateral oblique view of a left breast that contains a cluster of microcalcifications associated with an early malignancy in the upper portion of the image. Our computer algorithm detected the true cluster but also called three false-positive clusters. Figure 5 shows a mediolateral oblique view of a left breast containing a cluster of benign microcalcifications in the mid upper portion. The computer detected the true cluster as well as one false-positive cluster.

Figure 6, which compares the composite ROC curves obtained in the three reading conditions, shows that radiologists' performance was poorest when they read the mammograms without CAD ($A_z = 0.94$). With level 1 CAD, the ROC curve is improved ($A_z = 0.97$), and with level 2 CAD, the ROC curve improves further ($A_z = 0.98$). The difference between the "without CAD" and the "level 1 CAD" curves is statistically significant at $P < 0.001$; the difference between the "without CAD" and "level 2 CAD" curves is also statistically significant at $P < 0.001$. The difference between the "level 1 CAD" and "level 2 CAD" curves falls short of statistical significance ($P = 0.24$).

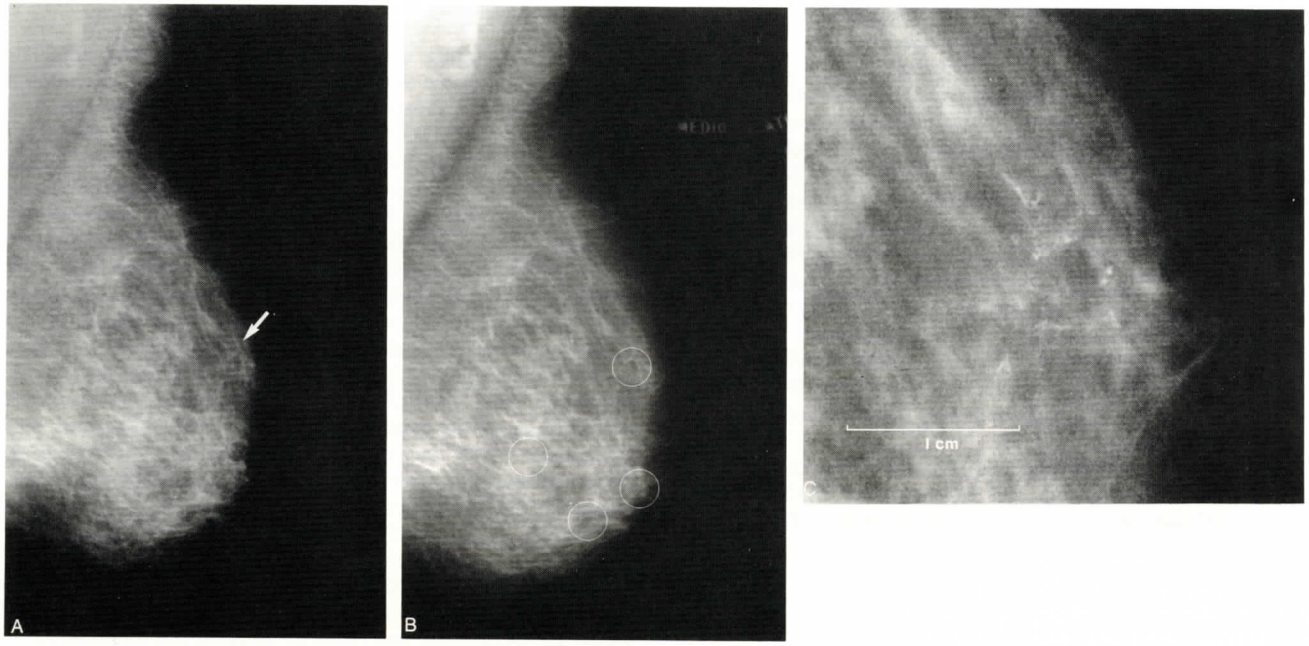

Fig. 4. A paired display (A and B) of an abnormal mammogram used in the ROC study. A cluster of microcalcifications associated with malignancy is present in the upper portion on a mediolateral oblique view of the left breast. (A) Image without computer output for primary reading. (B) Image superimposed with locations of microcalcification clusters detected by the computer at level 1 accuracy. The computer detected the true cluster and three FP clusters. (C) A close-up view of the subtle cluster of microcalcifications. The scale corresponds to the size on the image printed by the laser printer.

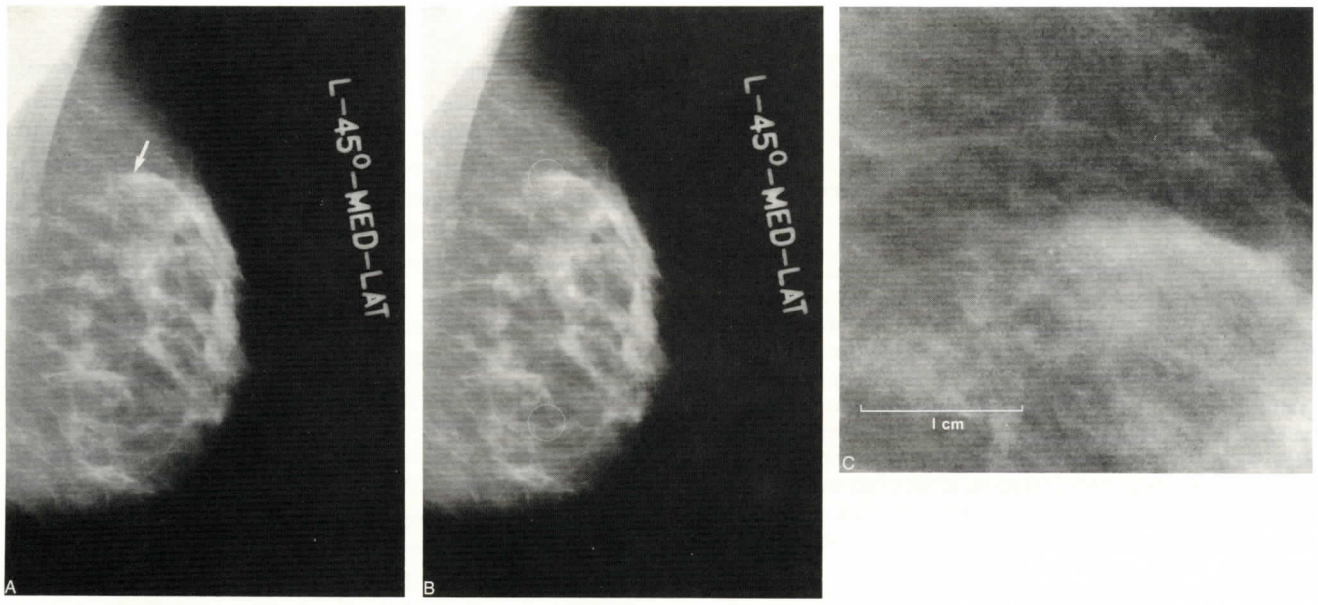

Fig. 5. A paired display (A and B) of an abnormal mammogram used in the ROC study. A cluster of benign microcalcifications is present in the mid-upper portion on a mediolateral oblique view of the left breast. (A) Image without computer output for primary reading. (B) Image superimposed with locations of microcalcification clusters detected by the computer at level 1 accuracy. The computer detected the true cluster and one FP cluster. (C) A close-up view of the subtle cluster of microcalcifications. The scale corresponds to the size on the image printed by the laser printer.

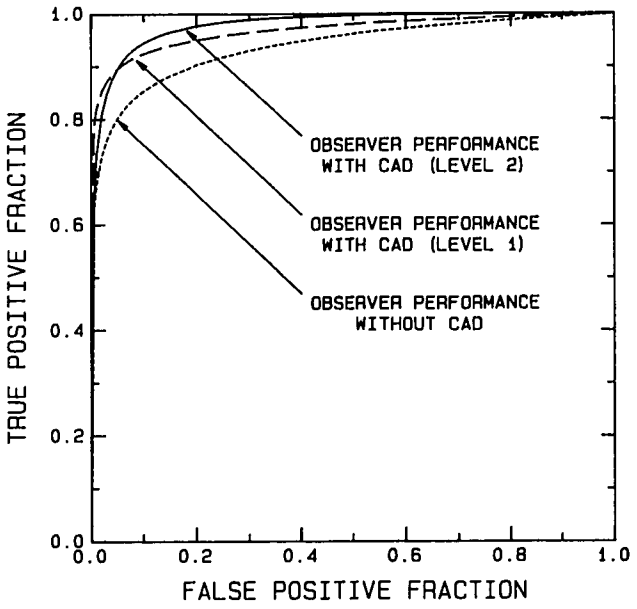

Fig. 6. Composite ROC curves for the three reading conditions obtained from the observer performance study.

Pooled LROC curves for the three reading conditions are compared in Figure 7. The relative ranking of the three conditions remains the same as that determined by the conventional ROC curves. However, with localization, the LROC curves indicate more clearly that the accuracy of radiologists in identifying the true clustered microcalcifications increases with CAD. At each confidence level, the sensitivity of cluster detection increases without an appreciable compromise in specificity (1-FPF). The sensitivity of the observers appears to increase if the computer detects fewer FP clusters (level 2 CAD).

We also analyzed separately the ROC and LROC curves obtained from the seven attending radiologists and from the eight radiology residents, in an attempt to determine whether the effect of CAD on an observer's performance is dependent on the observer's clinical experience. For each subgroup of observers, the relative rankings of the composite ROC curves and the pooled LROC curves of the three reading conditions are the same as those shown in Figures 6 and 7, but because of the smaller number of observers in each subgroup, the statistical significance of the differences between the ROC curves is not as high.

Discussion and Conclusions

A conventional ROC curve for the ability of the computer alone to detect clustered microcalcifications can be derived either directly from the computer-output results or indirectly from the FROC curve (Fig. 3) by means of the Bunch transform.[20] The two ROC curves thus derived for our data set are similar; the former is shown in Figure 8. When this curve is compared with the observers' ROC curve without CAD in Figure 6, it is evident that the detection accuracy of the current computer program alone is lower than that achieved by an average radiologist alone. This is attributable to the computer's inability to effectively discriminate false signals from true signals (Fig. 3). However, the results of our observer performance study indicate that when a computer program having high sensitivity is combined with a human observer's ability to rule out a majority of the

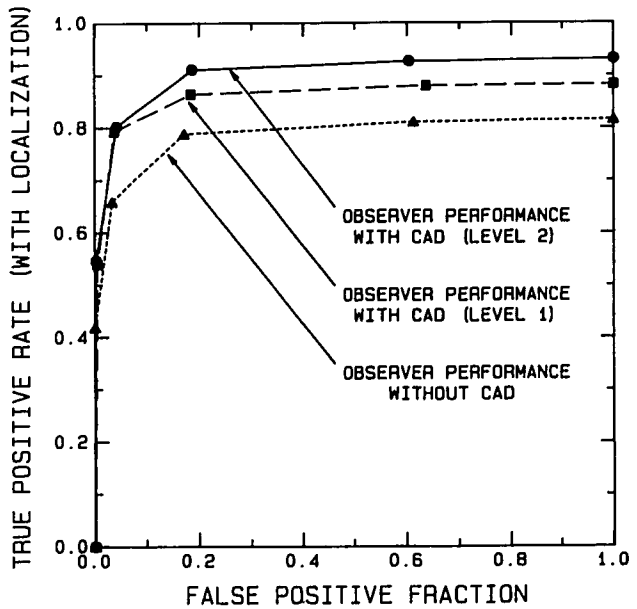

Fig. 7. Pooled LROC curves for the three reading conditions obtained from the observer performance study.

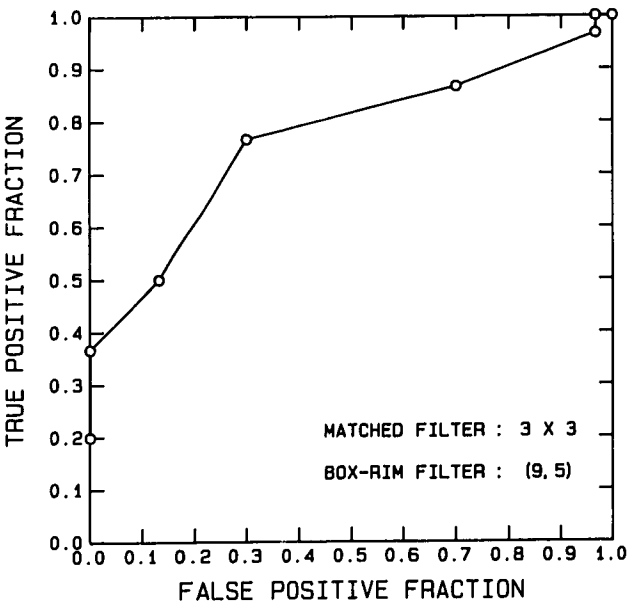

Fig. 8. ROC curve achieved by the computer detection program alone for the 60 clinical mammograms.

computer's FP detections, an overall improvement in performance can be realized. The small statistically insignificant improvement in the observers' ROC curve in moving from level 1 CAD to level 2 CAD (Fig. 6) suggests that radiologists' performance is not strongly influenced by a moderate number of FP detections by the computer in a CAD scheme. Further, because some of the true clusters detected by the computer alone may be missed by a human observer alone and vice versa, the overall detectability of TP clusters is higher when a human observer works with CAD than when the computer or the human observer works alone (compare Figures 6 and 8).

Table 1 tabulates the areas A_z under the individual observers' ROC curves and the differences between pairs of the observer-specific A_z values. Degenerate ROC data[22] were excluded from averages and from calculations of the composite ROC curves. The standard deviation of the A_z values for each reading condition shown in Table 1 indicates interobserver variability, which is quite consistent across the three reading conditions. The differences between the paired data reveal an improvement in detection accuracy with CAD for a majority of the observers. Thus, it appears that CAD can improve the performance of both experienced radiologists and radiology residents in the detection of subtle microcalcifications.

We compared the confidence ratings, averaged over the 15 observers, for each mammogram read without and with CAD. The results are listed in Table 2. The confidence ratings produced by abnormal cases generally increase with CAD, whereas for the normal cases, the confidence ratings remain essentially unchanged. These results from the ROC study agree with subjective impressions expressed by the observers after the experiment. In effect, the computer output seems to have served as a "second opinion." If the computer detected a subtle cluster of microcalcifications that was also seen by the observer, then that reassurance appears to have increased the observer's confidence in a positive interpretation to a level higher than that obtained from the mammogram alone.

TABLE 2. Numbers of Abnormal and Normal Cases for Which CAD Changed Observers' Average Confidence Ratings

	30 AbnormalCases		30 Normal Cases	
CAD	Level 1	Level 2	Level 1	Level 2
Increase	23	24	14	12
No change	1	2	4	4
Decrease	6	4	12	14

We also analyzed the number of observers who correctly identified the location of the microcalcification cluster on each abnormal mammogram with and without CAD. The results are shown in Table 3. For most images, correct localization increases with CAD, in agreement with the overall results shown by the pooled LROC curves (Fig. 7). Interestingly, the five cases for which correct localization decreases with level 1 CAD include all four cases in which computer detection was falsely negative; similarly, all three such cases with level 2 CAD were falsely read as negative by the computer. These results have two implications. First, although the computer program can detect many subtle clustered microcalcifications that may potentially be missed by radiologists, the program in its present state of development still fails to detect some of the very subtle clusters that are likely to be missed also by a radiologist. Improvement in the sensitivity of the computer program is therefore needed. Second, in our experimental setting, an observer who did not immediately see any microcalcification cluster and was

TABLE 1. The Area (A_z) Under the Individual Observer's ROC Curves and the Difference Between Pairs of A_z Values for the Three Reading Conditions

	A_z			Difference of A_z		
Observer	No CAD	L1 CAD	L2 CAD	L1-No	L2-No	L1-L2
1	0.953	0.989	0.975	0.036	0.022	−0.014
2	0.957	0.989	0.987	0.032	0.030	−0.002
3	0.952	0.992	0.993	0.040	0.040	0.000
4	0.931	0.949	0.993	0.018	0.061	0.044
5	0.941	—	0.940	—	−0.001	—
6	0.939	0.952	0.973	0.014	0.034	0.021
7	0.947	0.940	0.983	−0.007	0.036	0.043
8	0.920	0.973	0.949	0.053	0.029	−0.024
9	0.867	0.900	0.980	0.033	0.113	0.080
10	0.836	0.860	0.913	0.024	0.077	0.053
11	0.959	0.972	0.991	0.013	0.032	0.019
12	0.983	—	0.985	—	0.002	—
13	0.940	0.991	—	0.051	—	—
14	0.896	0.974	0.984	0.079	0.089	0.010
15	0.840	0.911	0.846	0.072	0.007	−0.065
Mean	0.924	0.953	0.964	0.035	0.041	0.014
SD	0.044	0.041	0.041	0.024	0.033	0.039
SD of mean				0.007	0.009	0.011

The blank entries indicate degenerate ROC data.[22]
L1: level 2; L2: level 1

TABLE 3. Number of Abnormal Cases that Showed a Change, with CAD, in the Number of Observers Who Correctly Identified the Location of the Microcalcification Cluster

CAD	Level 1	Level 2
Increase	17	20
No change	8	7
Decrease	5	3

"reassured" by the computer that the case was negative sometimes may have given up the search prematurely, thus resulting in a greater chance that a subtle cluster would be missed. The way in which this situation would transfer to clinical settings is likely to depend on factors such as the method used to present the computer's results to the radiologist, and the radiologist's understanding of the capability and limitations of the computer program. These factors must be more fully investigated and understood before CAD can be implemented effectively in clinical settings.

A limit of 5 seconds was imposed on the viewing time for each image or image pair in the observer experiments. It has recently[27] been suggested that the time needed for screening a four-view mammographic study for significant pathology can be as short as 45 seconds. Because in clinical practice a radiologist must search each of the images for all types of abnormalities, as well as evaluate comparable views of the right and left breast for asymmetries, we chose to limit the viewing time for microcalcifications on a single film to approximately one half the total viewing time expected for a single film, using 45 seconds as a guide. This approximated a clinical setting in which a high volume of films would be reviewed (eg, a screening program). It is in such situations that observer oversight might be expected to play a significant role in missed diagnoses.

Although two films were displayed with CAD, the observer did not need to spend additional time reading the second image because the underlying diagnostic information in each was identical. Also, it is reasonable to expect that one requisite for general acceptance of a CAD method will be that it does not increase reading time considerably. In our experimental setting, the observers who read the image without circles first and then checked the circled locations for microcalcifications thought that CAD would not increase their reading time. Further, the observers who read the circled locations first believed that CAD reduced their reading time. During the experiment, it was evident that most of the ratings given by the observers were recorded before the 5-second limit had been reached. Therefore, we believe that limiting viewing time to 5 seconds did not affect the relative ranking of the three reading conditions because of insufficient reading time for the paired images, and we expect that CAD would not reduce radiologists' efficiency in reading mammograms in actual clinical situations.

In the ROC study, we compared observer performance in the detection of subtle clustered microcalcifications with and without CAD. All mammograms used in the study were digitized with a 0.1 mm × 0.1 mm pixel size and were printed with a laser printer having somewhat lower spatial resolution. The lower spatial resolution of the reconstituted images caused the microcalcifications to appear to be slightly more subtle than those on the original mammograms. However, by visual comparison of the original and reconstituted images, we found that the degradation in the visibility of the microcalcifications was much less noticeable than that observed for the case samples used in our previous study,[28] probably because the microcalcifications were not as extremely subtle as those included in the previous cases. Because the same set of digitized films was used in all three reading conditions, we expect that the same relative ranking of the three reading conditions would be found with higher-spatial-resolution images, such as original screen-film mammograms, but with case samples that contain microcalcifications of a degree of subtlety corresponding to those on the reconstituted images. It should be noted that the purpose of our study is not to suggest replacing interpretation of screen-film mammograms with interpretation of digitized mammograms with CAD, nor to quantify the absolute amount of improvement or the absolute detection accuracy of the observers for a particular set of case samples. Such results will be of limited value because they will depend strongly on factors such as the degree of subtlety of the case samples, the accuracy of the computer program at the time of implementation, and the experience of the radiologists. Rather, we investigated the effects of CAD on mammographic interpretation and thus provided information for future development of CAD schemes. Our important finding is that a properly designed and implemented CAD scheme can provide statistically significant improvement in detection accuracy when the lesions are sufficiently subtle that they may not all be detected by the radiologists alone. This result should be applicable to interpretation of either conventional screen-film mammograms or primary digital mammograms that may be acquired directly for clinical examination in the future.

Acknowledgments

This work is supported by USPHS grant CA 48129, a Faculty Research Award (FRA-334) by the American Cancer Society, and USPHS grant CA 24806.

The authors are grateful to Drs. M. H. Awh, M. Backus, R. J. Foust, P. M. Jokich, C. E. Kahn, Jr., D. N. Levin, V. Martich, S. M. Montner, A. C. Philbrick, Y. Sasaki, C. A. Sennett, A. E. Stillman, and E. Wong for their participation in the observer performance studies; to H. B. Wang and M. Carlin for their technical assistance; and to E. Ruzich for secretarial assistance.

References

1. Martin JE, Moskowitz M, Milbrath JR. Breast cancer missed by mammography. AJR 1979;132:737–739.
2. Vernon MD. The psychology of perception. Middlesex, England: Penguin; 1962.

3. Tuddenham WJ. Visual search, image organization and reader error in roentgen diagnosis. Radiology 1962;78:694–704.
4. Smith MJ. Error and variation in diagnostic radiology. Springfield, IL: Charles C. Thomas Publisher; 1967.
5. Lehr JL, Lodwick GS, Farrell C, Braaten MO, Virtama P, Koivisto EL. Direct measurement of the effect of film miniaturization on diagnostic accuracy. Radiology 1976;118:257–263.
6. Hillman BJ, Fajardo LL, Hunter TB, et al. Mammogram interpretation by physician assistants. AJR 1987;149:907–911.
7. Stitik FP, Tockman MS. Radiographic screening in the early detection of lung cancer. Radiol Clin North Am 1978;16:347–366.
8. Yerushalmy J. The statistical assessment of the variability in observer perception and description of roentgenographic pulmonary shadows. Radiol Clin North Am 1969;7:381–392.
9. Guiss LW, Kuensler P. A retrospective view of survey photofluorograms of persons with lung cancer. Cancer 1960;13:91–95.
10. Hessel SJ, Herman PG, Swensson RG. Improving performance by multiple interpretations of chest radiographs: effectiveness and cost. Radiology 1978;127:589–594.
11. Kalisher L. Factors influencing false negative rates in xeromammography. Radiology 1979;133:297–301.
12. Bassett LW, Bunnell DH, Jahanshahi R, Gold RH, Arndt RD, Linsman J. Breast cancer detection: One versus two views. Radiology 1987;165:95–97.
13. Moskowitz M: Benefit and risk. In: Bassett LW, Gold RH, eds. Breast cancer detection: Mammography and other methods in breast imaging. ed. 2. New York, NY: Grune and Stratton; 1987;131–142.
14. Baines CJ, Miller AB, Wall C, et al. Sensitivity and specificity of first screen mammography in the Canadian National Breast Screening Study: A preliminary report from five centers. Radiology 1986; 160:295–298.
15. Haug PJ, Tocino IM, Clayton PD, Bair TL. Automated management of screening and diagnostic mammography. Radiology 1987; 164:747–752.
16. Chan HP, Doi K, Galhotra S, Vyborny CJ, MacMahon H, Jokich PM. Image feature analysis and computer-aided diagnosis in digital radiography: automated detection of microcalcifications in mammography. Med Phys 1987;14:538–548.
17. Chan HP, Doi K, Vyborny CJ, Lam KL, Schmidt RA. Computer-aided detection of microcalcifications in mammograms: methodology and preliminary clinical study. Invest Radiol 1988;23:664–671.
18. Ishida M, Kato H, Doi K, Frank PH. Development of a new digital radiographic image processing system. Proc SPIE 1982;347:42–48.
19. Katsuragawa S, Doi K, MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography: detection and characterization of interstitial lung disease in digital chest radiographs. Med Phys 1988;15:311–319.
20. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A free response approach to the measurement and characterization of radiographic observer performance. Proc SPIE 1977;127:124–135.
21. MacMahon H, Metz CE, Doi K, Kim T, Giger ML, Chan HP. Digital chest radiography: Effect on diagnostic accuracy of hard copy, conventional video, and reversed gray scale video display formats. Radiology 1988;168:669–673.
22. Metz CE. Practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989;24:234–245.
23. Swets JA, Pickett RM. Evaluation of diagnostic systems: Methods from signal detection theory. New York, NY: Academic Press; 1982.
24. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986;21:720–733.
25. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. J Math Psychol 1969;6:487–496.
26. Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. Radiology 1975;116:533–538.
27. Sickles EA, Weber WN, Galvin HB, Ominsky SH, Sollitto RA. Mammographic screening: how to operate successfully at low cost. Radiology 1986;160:95–97.
28. Chan HP, Vyborny CJ, MacMahon H, Metz CE, Doi K, Sickles EA. Digital mammography: ROC studies of the effects of pixel size and unsharp mask filtering on the detection of subtle microcalcifications. Invest Radiol 1987;22:581–589.