

UNCLASSIFIED

AD NUMBER
ADB274470
NEW LIMITATION CHANGE
TO Approved for public release, distribution unlimited
FROM Distribution authorized to U.S. Gov't. agencies only; Proprietary information, Aug 2001. Other requests shall be referred to Army Medical Research and Materiel Command, 504 Scott St., Fort Detrick, MD 21702-5012.
AUTHORITY
USAMRMC ltr, 26 Aug 2002

THIS PAGE IS UNCLASSIFIED

AD \_\_\_\_\_

Award Number: DAMD17-98-1-8018

TITLE: False-Negative Interpretations in a CAD Environment

PRINCIPAL INVESTIGATOR: Bin Zheng, Ph.D.

CONTRACTING ORGANIZATION: University of Pittsburgh  
Pittsburgh, Pennsylvania 15260

REPORT DATE: August 2001

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command  
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Distribution authorized to U.S. Government agencies only (proprietary information, Aug 01). Other requests for this document shall be referred to U.S. Army Medical Research and Materiel Command, 504 Scott Street, Fort Detrick, Maryland 21702-5012.

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20020124 311

## NOTICE

USING GOVERNMENT DRAWINGS, SPECIFICATIONS, OR OTHER DATA INCLUDED IN THIS DOCUMENT FOR ANY PURPOSE OTHER THAN GOVERNMENT PROCUREMENT DOES NOT IN ANY WAY OBLIGATE THE U.S. GOVERNMENT. THE FACT THAT THE GOVERNMENT FORMULATED OR SUPPLIED THE DRAWINGS, SPECIFICATIONS, OR OTHER DATA DOES NOT LICENSE THE HOLDER OR ANY OTHER PERSON OR CORPORATION; OR CONVEY ANY RIGHTS OR PERMISSION TO MANUFACTURE, USE, OR SELL ANY PATENTED INVENTION THAT MAY RELATE TO THEM.

1

### LIMITED RIGHTS LEGEND

Award Number: DAMD17-98-1-8018  
Organization: University of Pittsburgh

Those portions of the technical data contained in this report marked as limited rights data shall not, without the written permission of the above contractor, be (a) released or disclosed outside the government, (b) used by the Government for manufacture or, in the case of computer software documentation, for preparing the same or similar computer software, or (c) used by a party other than the Government, except that the Government may release or disclose technical data to persons outside the Government, or permit the use of technical data by such persons, if (i) such release, disclosure, or use is necessary for emergency repair or overhaul or (ii) is a release or disclosure of technical data (other than detailed manufacturing or process data) to, or use of such data by, a foreign government that is in the interest of the Government and is required for evaluational or informational purposes, provided in either case that such release, disclosure or use is made subject to a prohibition that the person to whom the data is released or disclosed may not further use, release or disclose such data, and the contractor or subcontractor or subcontractor asserting the restriction is notified of such release, disclosure or use. This legend, together with the indications of the portions of this data which are subject to such limitations, shall be included on any reproduction hereof which includes any part of the portions subject to such limitations.

THIS TECHNICAL REPORT HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION.

  
\_\_\_\_\_

12/25/71  
  
\_\_\_\_\_

# REPORT DOCUMENTATION PAGE

Form Approved  
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

<b>1. AGENCY USE ONLY (Leave blank)</b>		<b>2. REPORT DATE</b> August 2001	<b>3. REPORT TYPE AND DATES COVERED</b> Final (1 Jul 98 - 1 Jul 01)	
<b>4. TITLE AND SUBTITLE</b> False-Negative Interpretations in a CAD Environment			<b>5. FUNDING NUMBERS</b> DAMD17-98-1-8018	
<b>6. AUTHOR(S)</b> Bin Zheng, Ph.D.				
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b> University of Pittsburgh Pittsburgh, Pennsylvania 15260  Email: bzheng@radserv.arad.upmc.edu			<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b>	
<b>9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b> U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012			<b>10. SPONSORING / MONITORING AGENCY REPORT NUMBER</b>	
<b>11. SUPPLEMENTARY NOTES</b>				
<b>12a. DISTRIBUTION / AVAILABILITY STATEMENT</b> Distribution authorized to U.S. Government agencies only (proprietary information, Aug 01). Other requests for this document shall be referred to U.S. Army Medical Research and Materiel Command, 504 Scott Street, Fort Detrick, Maryland 21702-5012.			<b>12b. DISTRIBUTION CODE</b>	
<b>13. ABSTRACT (Maximum 200 Words)</b>  The purpose of this project is to examine the impact of CAD schemes on the diagnostic performance of radiologists, especially the change of false-negative interpretations under a CAD cueing environment. Using an observer performance study, we hope to better understand the impact of CAD schemes on diagnostic performance for detection of subtle abnormalities and to find an optimal approach to use CAD schemes in a screening environment. In this project, seven radiologists participated in the study. Each interpreted 120 subtle mammographic cases in five different reading modes (including a non-cued mode and four CAD-cued modes). The study results demonstrated that CAD cueing with high accuracy (i.e., 90% sensitivity and 0.5 false-positive per image) significantly improved observer's performance ( $p < 0.05$ ). As accuracy of CAD cueing decreased so did observer performance. Reducing cueing sensitivity and specificity increased false-negative rates in non-cued areas ( $p < 0.05$ ). In conclusion, this study indicates that in a laboratory environment, observer performance in the detection of subtle mammographic abnormalities is significantly affected by the inherent performance of a CAD cueing system.				
<b>14. SUBJECT TERMS</b> Computer-assisted detection, false-negative interpretations, observer performance study, ROC analysis			<b>15. NUMBER OF PAGES</b> 51	
			<b>16. PRICE CODE</b>	
<b>17. SECURITY CLASSIFICATION OF REPORT</b> Unclassified	<b>18. SECURITY CLASSIFICATION OF THIS PAGE</b> Unclassified	<b>19. SECURITY CLASSIFICATION OF ABSTRACT</b> Unclassified	<b>20. LIMITATION OF ABSTRACT</b> Unlimited	

## Table of Contents

Cover.....	
SF 298.....	2
Table of Contents.....	3
Introduction.....	4
Body.....	4
Key Research Accomplishments.....	8
Reportable Outcomes.....	8
Conclusions.....	9
References.....	9
Project Personnel .....	10
Appendices.....	11

## INTRODUCTION

During the last decade, interest in computer-assisted diagnosis (CAD) schemes for the early detection of breast cancer on mammograms has been rapidly increasing, and a large variety of schemes have been developed and tested. As a result, it is believed that eventually CAD schemes could provide radiologists with useful information to improve the efficiency and accuracy in the diagnosis of breast cancer [1-4]. However, prompting potential areas of abnormalities can affect the mammographic interpretation process, and unfortunately, the effect may not be always beneficial [5-8]. Therefore, to better understand the radiological interpretation process, we conducted an experiment to examine how different CAD cueing environments affect the error rate (particularly for false-negative interpretations). For this purpose, we first selected a set of subtle cases (including the mix of subtle malignant and difficult negative cases). These cases were cued under different levels of sensitivity and specificity based on a CAD scheme's processing. Radiologists were then selected to participate in an observer performance experiment to view and score these cases under five different reading modes (including one non-cued and four CAD-cued modes). From the experimental results, we performed different analyses using ROC-type methodology. From the relationship between the CAD cueing levels and average diagnostic performance, we hope not only to better understand the impact of CAD cueing on diagnostic performance, but also evaluate an optimal approach to use CAD schemes in the screening environment.

## BODY – SUMMARY OF WORK PERFORMED

In the past three years of this project, we performed and completed the following tasks:

### Year 1:

#### 1. Case selection:

From a large pool of initial cases, using a comprehensive case verification, categorization, and selection protocol, we selected 120 subtle mammographic examinations from 120 patients undergoing routine examinations in three different medical centers. In these 120 cases, 85 were abnormal and 35 were negative. The abnormal cases involved a total of 38 verified microcalcification clusters (27 malignant and 11 benign) and 57 masses (39 malignant and 18 benign). Most of these cases involve two images (the same view of left and right breasts), but in some, we used only one image. Table 1 summarizes the number of cases in different categories. All positive cases were verified using source documents. All the negative cases were determined based on the current and follow-up mammographic examination results. All the cases were considered to be "subtle" by radiologists, because these involve either subtle abnormalities or complex normal anatomy. Original film mammograms were digitized in our laboratory using

the same high quality film digitizer with resolution of 12-bit gray levels and 100  $\mu\text{m} \times 100 \mu\text{m}$  pixel sizes.

Table 1: Distribution of cases in different categories. (M – malignant, B – benign).

	Mass		Microcalcification Cluster		Mass and Cluster		Negative	Total Cases
	M	B	M	B	M	B		
Single image case	10	1	11	3	1	1	4	31
Two image cases	20	16	7	7	8	0	31	89
Total Cases	30	17	18	10	9	1	35	120

## 2. CAD processing and cueing mode design:

In this study one non-cued mode and four CAD cued-modes were designed (as shown in Table 2). The four cued modes emulated what can be expected using current levels of CAD performance as well as what one hopes to achieve using CAD in the future. To generate the cues (including masses and microcalcification clusters), every image in the selected database was first processed by our own CAD schemes [9-11]. Each suspicious region detected by the CAD schemes received a likelihood score for being positive (from 0 to 1). The larger the score, the more “likely” the region was estimated to represent a true-positive region. Based on the detection scores, we separately selected cueing regions for true-positive regions and false-positive regions in different cueing modes to meet the cueing requirements as listed in Table 2.

Table 2: Five reading modes in the experiment.

Reading mode	ROI Marked	Marked sensitivity	Marked FP / image
1	No	0	0
2	Yes	0.9	0.5
3	Yes	0.9	2
4	Yes	0.5	0.5
5	Yes	0.5	2

### 3. Implementation of a computer-controlled image display system.

The reading in this study was performed directly from the digitized images displayed on a SUN-workstation monitor. Each reading session was designed to include 30 cases. Hence, each observer had 20 reading sessions (by reading 120 cases using five different display modes) in the study. To prepare such a reading experiment, we designed, tested, and implemented an automatic image display and control system. First, we assigned a file containing a counterbalanced order of reading modes for each observer. The 20 sessions were divided into 4 blocks with 5 sessions each. In each block, one observer read five sessions with five different modes in a random order. Second, the computer 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 for each observer were different. A minimum time delay between two consecutive readings of the same cases was also imposed in the program. Third, a computer display and control program was designed and implemented. The images can be alternatively displayed at two display modes. The observer can view two images side by side displayed on the monitor at a reduced resolution to fit monitor size. By changing display mode, the observer can also examine full resolution images, one at a time using scrollbars in both vertical and horizontal directions (zoom and scroll). A "Display/Remove" button could be used to superimpose or delete the CAD cues on the images. An observer could make a diagnostic decision while viewing either sub-sampled or full-resolution images. A management computer program was designed to automatically record all diagnostic information entered by the observers, including the type of a detected abnormality (mass or microcalcification cluster), its location (the center of the detected region), and two estimated likelihood scores for the detection (presence/absence) and classification (benign/malignant) of "detected" regions.

### 4. Finalizing study protocol.

We have thoroughly tested the reliability of our computer management system in our laboratory for this study and have written a comprehensive set of "Instructions for the Readers." The instruction provided the detailed requirement of the reading experiments and described how to use the computer-controlled image display and scoring system.

### Year 2:

### 5. Selection of readers and pre-training

In the second year of the project, after we finalized the study protocol, we set up our soft-display workstation in a clinical reading room of our medical center and recruited radiologists to participate in this study. All the radiologists selected were Board certified radiologists with a minimum of three years' experience in the interpretation of mammograms. Before the main study, we first provided each reader a comprehensive written "Instructions for the Readers." Then, each reader had a training session using a set of sample cases. The written instruction and

the training session enabled each reader to get familiar with the reading environment (soft-display image and diagnostic scoring system) in the main reading experiment. We originally selected eight radiologists to participate in the study, but the blind nature of the study precluded us from recognizing that one of them completely misinterpreted the task at hand and reported the cases under a totally different set of rules, which was only found at the end of the study. Although this reader's results led to the same conclusions, the "rules" under which he/she reported cases were so different that we excluded the results from the analysis, and a separate analysis is underway and will result in a separate publication.

#### 6. Performing the main reading experiment.

The main reading experiment was carried throughout the second year at our medical center. Each reader interpreted 120 cases five times (five display modes) in a counterbalanced order. We divided the reading process into 20 reading sessions. Each reading session included 30 cases. To reduce the bias caused by the possible recognition of specific cases read previously, a minimum time delay (10 days) between two consecutive readings of the same case was implemented. By the end of the year, six observers completed all reading sessions. The remaining observers also completed the majority of their reading sessions.

#### Year 3:

#### 7. Completing the main reading experiment.

This observer performance experiment under different CAD cueing modes was completed in the third year of this project. The reading sessions were finished before November 1, 2000. All data collected during the entire reading period were then examined and transferred from the experimental workstation for analyses.

#### 8. Data analysis.

Although detection performance varied among the observers who participated in the study, we found that a general pattern of performance curves was consistent for all observers (with the exception of the one noted previously). Hence, the data from seven observers were pooled together in most of our data analyses. The reading results have been analyzed using FROC and other statistical methodologies (i.e., trend analysis using two-way analysis of variance (AOV) [12] and a Wilcoxon matched pair test). The ROCFIT program developed by Dr. Metz et al at the University of Chicago was used to generate the performance curves [13]. From these FROC curves generated for the different reading modes, we compared detection sensitivity at ten different false-positive rates uniformly distributed over the measured range. Both sensitivity and specificity of CAD cueing results affected observer performance. The trend was significant ( $p < 0.05$ ). The data analysis from Wilcoxon matched pair tests demonstrated that increasing false-positive cueing could significantly increase false-negative detections in both cued and non-cued areas ( $p < 0.05$ ). To assess the reliability of our data analysis procedure (or reduce potential bias), we conducted two additional tests. First, we re-ordered the reading results by grouping all cases that were read for the first time (regardless of mode) as one group, and all cases read for

the second time as another group, etc. The performance curves were computed separately for these five mutually exclusive groups and were compared. Second, we excluded all single-image cases from the database and compared the detection results on all cases with bilateral images. The analysis results from these two additional tests were consistent with our initial observations. Finally, we also analyzed the pooled classification ratings (malignant vs. benign) provided by these observers. The result showed that once identified (detected), observers' ability to distinguish between malignant vs. benign abnormalities (classification) were not affected ( $p > 0.05$ ) by the cueing mode or lack thereof [14].

### KEY RESEARCH ACCOMPLISHMENTS

- We performed a blind multi-abnormality, multi-mode, multi-observer performance study using a set of 120 subtle cases and a large number of experienced observers.
- We demonstrated for the first time that although CAD systems with high accuracy had the potential of significantly improving diagnostic performance in mammography, poorly performing schemes could adversely affect observer performance in both cued and non-cued areas.

### REPORTABLE OUTCOMES

The detailed results (including experimental design and statistical analysis of the experimental data) from this observer performance study have been reported in our manuscript, which has been accepted for publication by *Radiology*, as cited below:

- Zheng B, Ganott MA, Britton CA, Hakim CM, Hardesty LA, Chang TS, Rockette HE, Gur D, Soft-display mammographic readings under different computer-assisted detection cueing environments: Preliminary findings, *Radiology* 2001; accepted for publication

In addition, we also published two related CAD papers, which used a similar analytic method that we applied in this project and acknowledged the support of this research grant:

- 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, *Academic Radiology* 2000; 7:595-602.
- Zheng B, Chang YH, Good WF, Gur D, Performance gain in Computer-assisted detection schemes by averaging scores generated from artificial neural networks with adaptive filtering, *Medical Physics* 2001; (conditionally accepted / pending minor revision).

## CONCLUSION

There are five research tasks listed in the Statement of Work of this project. In the first year, we completed the first three tasks. Task four (main reading) was carried out through the second year and completed in the first several months of the third year. Task five (data analysis) was completed in the third year. In this project, a blind multi-abnormality, multi-mode, multi-observer performance study was performed to examine the impact of different CAD cueing modes on the diagnostic performance of radiologists. Such an observer performance study has not been performed and reported elsewhere.

### So What:

From this study we demonstrated that in a laboratory environment, observer performance in the detection of subtle mammographic abnormalities was significantly affected by the inherent performance of a cueing system. A good (or "highly" accurate) CAD cueing system (i.e., 90% cueing sensitivity and 0.5 false-positive cueing rate) could significantly improve diagnostic performance of radiologists, while a poorly performing CAD cueing system could actually degrade their diagnostic performance. Because this is a relatively small and preliminary study, large-scale studies are required to further address and confirm many issues discussed in this project. Although the experimental results and conclusions generated in this project are considered preliminary, this study clearly indicates the need for (1) appropriate usage of CAD cueing systems by radiologists and (2) further improvement of the accuracy of current CAD schemes.

## REFERENCES

1. Vyborny CJ, Giger ML, Computer vision and artificial intelligence in mammography, *AJR* **1994**; 162:699-708.
2. Hoffman KR, For 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**; 26:1-2.
3. 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.
4. Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA, Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection, *Radiology* **2001**; 219:192-202.
5. Krupinski EA, Nodine CF, Kundel HL, Perceptual enhancement of tumor targets in chest x-ray images, *Perception & Psychophysics* **1993**; 53:519-526.
6. Funovics M, Schamp S, Lackner B, Wolf G, Computer-assisted diagnosis in mammography: the R2 ImageCheck system in detection of speculated lesions, *Wien Med Wochenschr* **1998**; 3338:840-844.
7. 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.

8. Malich A, Azhari T, Bohm T, Fleck M, Kaiser WA, Reproducibility – an important factor determining the quality of computer aided detection (CAD) systems, *Euro Radiology* **2000**; 36:170-174.
9. 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.
10. 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.
11. 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.
12. Abelson PR, Tukey JW, Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order, *Ann of Mathematical Statistics* **1963**; 34:1347-1369.
13. Metz CE, Herman BA, Shen JH, Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data, *Stats in Med* **1998**; 17:1033-1053.
14. Zheng B, Ganott MA, Britton CA, Hakim CM, Hardesty LA, Chang TS, Rockette HE, Gur D, Soft-display mammographic readings under different computer-assisted detection cueing environments: Preliminary findings, *Radiology* **2001**; accepted for publication

#### PROJECT PERSONNEL

- John M. Drescher, B.S., Systems' Programmer
- Richard G. Swensson, Ph.D., Investigator / Statistician
- He Wang, Ph.D., Investigator / Data Analysis
- Xiao Hui Wang, Ph.D., Investigator / Case collection and preparation
- Bin Zheng, Ph.D., Principal Investigator

## APPENDICES

Two manuscripts are enclosed as appendices in this final report.

1. 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, *Academic Radiology* 2000; 7:595-602.
2. Zheng B, Ganott MA, Britton CA, Hakim CM, Hardesty LA, Chang TS, Rockette HE, Gur D, Soft-display mammographic readings under different computer-assisted detection cueing environments: Preliminary findings, *Radiology* 2001; accepted for publication

---

# Applying Computer-assisted Detection Schemes to Digitized Mammograms after JPEG Data Compression: An Assessment<sup>1</sup>

Bin Zheng, PhD, Jules H. Sumkin, DO, Walter F. Good, PhD, Glenn S. Maitz, MS  
Yuan-Hsiang Chang, PhD, David Gur, ScD

---

**Rationale and Objectives.** The authors' purpose was to assess the effects of Joint Photographic Experts Group (JPEG) image data compression on the performance of computer-assisted detection (CAD) schemes for the detection of masses and microcalcification clusters on digitized mammograms.

**Materials and Methods.** This study included 952 mammograms that were digitized and compressed with a JPEG-compatible image-compression scheme. A CAD scheme, previously developed in the authors' laboratory and optimized for noncompressed images, was applied to reconstructed images after compression at five levels. The performance was compared with that obtained with the original noncompressed digitized images.

**Results.** For mass detection, there were no significant differences in performance between noncompressed and compressed images for true-positive regions ( $P = .25$ ) or false-positive regions ( $P = .40$ ). In all six modes the scheme identified 80% of masses with less than one false-positive region per image. For the detection of microcalcification clusters, there was significant performance degradation ( $P < .001$ ) at all compression levels. Detection sensitivity was reduced by 4%–10% as compression ratios increased from 17:1 to 62:1. At the same time, the false-positive detection rate was increased by 91%–140%.

**Conclusion.** The JPEG algorithm did not adversely affect the performance of the CAD scheme for detecting masses, but it did significantly affect the detection of microcalcification clusters.

**Key Words.** Breast neoplasms, diagnosis; breast radiography; computers, diagnostic aid; data compression; images, storage and retrieval.

---

Mammography is the most commonly used method for early detection of breast cancer (1,2). With the current recommendations for annual screening of women over 40 years of age and with gradually increased compliance, the total number of mammograms obtained each year is increasing (3). To improve the efficiency and effectiveness of mammographic screening in general and in rural or

underserved areas in particular, there has been an increasing interest in telemammography (4,5).

Because high-resolution mammograms require large data sets of 10–40 Mbyte each, digital management of such images, including transmission, display, and archiving, quickly becomes a problem, especially when one is dealing with a large number of images. Hence, image-compression techniques have been explored. Many image-compression methods, such as full-frame discrete cosine transform (6), wavelet-type decomposition (7), and data compression compatible with Joint Photographic Experts Group (JPEG) scheme (8), have been applied successfully to digitized mammograms. At the same time, a large number of computer-assisted detection (CAD) schemes have been developed during the past decade to help detect masses and microcalcification clusters with digitized mammograms

---

Acad Radiol 2000; 7:595–602

<sup>1</sup>From A433 Scaife Hall, Radiological Imaging Division, Department of Radiology, University of Pittsburgh, 3550 Terrace St, Pittsburgh, PA 15261-0001. Received December 16, 1999; revision requested February 23, 2000; revision received March 13; accepted March 16. Supported in part by the National Cancer Institute under grants CA62800, CA77850, and CA82912 and the U.S. Army under contract DAMD17-98-1-8018. Address correspondence to B.Z.

©AUR, 2000

(9). Several large-scale studies have evaluated the performance of CAD schemes in clinical mammographic screening (10,11). Many investigators believe that CAD schemes can and eventually will provide radiologists with useful "second opinion" information to improve the diagnostic accuracy and efficiency of mammography (12,13). Integrating CAD schemes into telemammography systems could provide practical advantages and could make mammography more accessible, less expensive, and more accurate.

We routinely refer to our own CAD schemes as computer-assisted *detection* schemes. Other investigators use the term to refer to computer-assisted (or computer-aided) *diagnosis*, which includes detection tasks either alone (6,9–11) or in combination with classification (ie, determining the nature of the abnormality) (14–16). In this article, "CAD" refers to the detection tasks only.

The feasibility of combining CAD schemes with telemammography systems has not been fully investigated. Theoretically, a CAD scheme can be applied either to the original digitized images prior to data compression or to the reconstructed images at the receiving site. The first approach requires more computing power at the sending site and may slow down the complete process. The second approach allows for CAD schemes to be applied off-line, which does not affect data transmission and may be more flexible. A prerequisite for the second approach is that image compression must not impair the performance of the CAD scheme. One study, which involved 25 regions of interest, examined the effects of image compression on computerized detection of a single microcalcification. The results demonstrated that, to avoid marked degradation in performance, the compression ratio should be limited to 3.6:1 or less for the discrete cosine transform algorithm and 9.6:1 or less for the enhanced discrete cosine transform with the LPHC, or Laplacian pyramid hierarchical coding, method (6). In the current study, we investigated the relationship between a JPEG compression algorithm and the performance of a CAD scheme for both masses and clustered microcalcifications. In this article, we describe the experimental procedure in detail and present the results we obtained when applying the scheme to a set of 952 images.

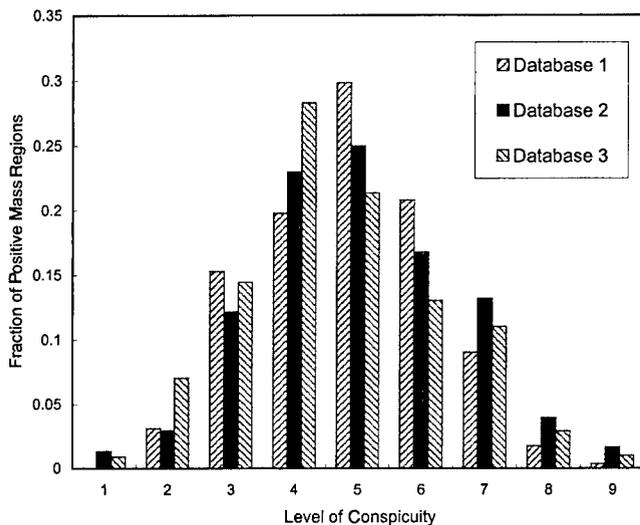
## MATERIALS AND METHODS

The image database selected for this study included 952 digitized mammograms. Among these, 424 were selected from a set of mammograms provided by a research group at Washington University, St Louis, Mo, and 528 were selected from images provided by the research and develop-

ment team at the Eastman Kodak Company, Rochester, NY. All mammograms were digitized in our laboratory; we used a laser-film digitizer (Lumisys, Sunnyvale, Calif) with a pixel size of  $100 \times 100 \mu\text{m}$  and 12-bit gray-level resolution. The quality of the digitizer was monitored routinely to ensure that gray levels (or pixel values) were linearly proportional to optical density in the range of 0.2–3.2. Our digitization protocol has been described in detail elsewhere (17). In addition to the image, the two groups provided the diagnostic information ("truth") for each case. This included the type of abnormalities (mass, microcalcification cluster, or both) and, when applicable, the location of the abnormalities and histopathologic results. On the 952 images, there were 408 regions depicting masses and 303 depicting microcalcification clusters. Biopsy results indicated that 264 regions depicted malignant masses, and 142 depicted microcalcification cluster regions associated with malignant cases.

Although we are not privy to the protocols used by the other groups for selecting the images that we used, we examined the feature distribution of these images and compared such features to those for the images collected at our own institution. We selected consecutively the cases with a diagnostic level of concern of at least 3. (With a rating scale system developed in our medical center in 1987, concern level 3 means "probably benign finding.") Using our own image database for comparison, we found a similar "detection difficulty" for the masses and microcalcification clusters identified on the images provided to us by the external groups. A larger proportion of the images provided by the research group at Eastman Kodak, however, contained dense parenchymal breast patterns. To describe detection difficulty, different measurements have been proposed, such as the effective size and contrast (18) or "visibility" as rated subjectively by radiologists (19). We typically use "conspicuity," defined as lesion contrast divided by surrounding complexity (20), to infer the difficulty of detecting mammographic abnormalities. The computing algorithm for conspicuity has been reported before (21). Figure 1 demonstrates the normalized distributions of conspicuity for positive mass regions in the three image databases. Lower conspicuity indicates greater difficulty in detecting the abnormality visually (22) and routinely with CAD schemes, as well (21).

A JPEG image-compression scheme has been previously developed and evaluated in our laboratory (8). In brief, each image was compressed to five different levels. Because in the JPEG algorithm it is the degree of quantization rather than the compression ratio itself that determines the



**Figure 1.** Normalized distributions of conspicuity for positive mass regions in the three image databases. Database 1 included our own cases; database 2, cases provided by Washington University; and database 3, cases provided by the research group at Eastman Kodak.

degradation of an image by compression, the quantization factor was used as an independent variable, and not the actual compression ratio. The five quantization factors we used are 40, 60, 80, 100, and 120, which produced average compression ratios of 17:1, 26:1, 35:1, 45:1, and 62:1 for this set of images. Each image is then decompressed (or reconstructed) into a restored image with the original matrix size.

All six images (one noncompressed and five compressed) for each case were processed by using a CAD scheme developed and tested in our laboratory (23,24). The CAD scheme includes two independent algorithms, one for detecting masses and one for detecting microcalcification clusters. For mass detection, every image is subsampled (pixel averaged) by a factor of four in both dimensions to reduce the size of each image to approximately  $600 \times 450$  pixels.

The algorithm for mass detection includes three distinct stages (24). The first stage of image segmentation (including dual kernel filtering, subtraction, thresholding, and labeling) is used to search for all regions in which masses are suspected. Depending on the complexity of breast tissue structure, approximately 10–30 regions are likely to be identified for each image. Then, based on local contrast measurement, an adaptive region growth algorithm to define three topographic layers is applied to each suspicious region. For each growth layer, a set of simple intralayer boundary conditions on growth ratio and shape factor of the region is applied to eliminate a large portion of suspi-

cious regions (about 70%–80%). For each of the remaining regions, a set of features (or a feature vector) is automatically computed. In the third stage, a nonlinear multilayer multifeature analysis is applied to classify suspected regions as positive or negative. In this study, a pretrained artificial neural network (ANN) (24) was incorporated into the CAD algorithm for mass detection. The ANN involves 16 features, which were selected by means of a genetic algorithm-based optimization process (25).

To detect microcalcification clusters, the following steps are applied. First, Gaussian band-pass filtering and local contrast-based thresholding are applied. A large number of suspicious pixels are identified in this step. Second, a special local minimum search ring is applied to analyze all selected pixels (26), followed by "blob" labeling and clustering. A topographic growth and rule-based feature analysis is used for each suspected microcalcification. The remaining suspected microcalcifications are clustered, and a set of rules related to the clustering is applied (26). The performance and robustness of this algorithm have been reported (27). Recently, an ANN classifier has been trained independently and replaced the last step of the algorithm. The ANN involves 13 input neurons, seven hidden neurons, and one output neuron. The definitions and computation methods of these features have been described elsewhere (26).

It should be emphasized that all of the decision rules and the two ANNs were trained independently with a different image database (24,27). This CAD scheme was applied without any modification ("as is") to the database tested here. Because images were compressed at five levels and reconstructed, the scheme was applied six times to each image. The detection results for all six testing modes were tabulated and compared.

The number of true-positive regions identified, the proportion of images without any false-positive regions, and the average number of false-positive regions per image were the summary indexes used to compare the results before the application of the ANNs. Because the number of images was relatively large, we assumed that the average value of the summary statistics was approximately normally distributed (the central limit theorem). For testing the hypothesis of equality in the proportion of images without any false-positive regions and the average number of images with false-positive findings across the six presentations, we used a modification of a test proposed by Abelson and Tukey (28) to identify trends within one-way analysis of variance. Tests for trends were one sided, with the hypothesized alternative showing a decrease in true-positive detections with increasing compression, or an increase in

**Table 1**  
**True-Positive and False-Positive Detection of Masses**

Finding	Compression Ratio					
	0	17:1	26:1	35:1	45:1	62:1
No. of true-positive mass regions	383	382	381	381	382	382
Average no. of false-positive regions per image	3.4	3.4	3.4	3.4	3.4	3.4
Percentage of images without any false-positive region	8	8	9	9	9	8

Note.—Regions with suspected masses identified with CAD before application of the ANN (16).

**Table 2**  
**True-Positive and False-Positive Detection of Microcalcification Clusters**

Finding	Compression Ratio					
	0	17:1	26:1	35:1	45:1	62:1
No. of true-positive cluster regions	298	285	285	284	269	270
Average no. of false-positive regions per image	2.6	5.0	5.4	5.2	5.5	6.3
Percentage of images without any false-positive region	35	18	16	14	14	10

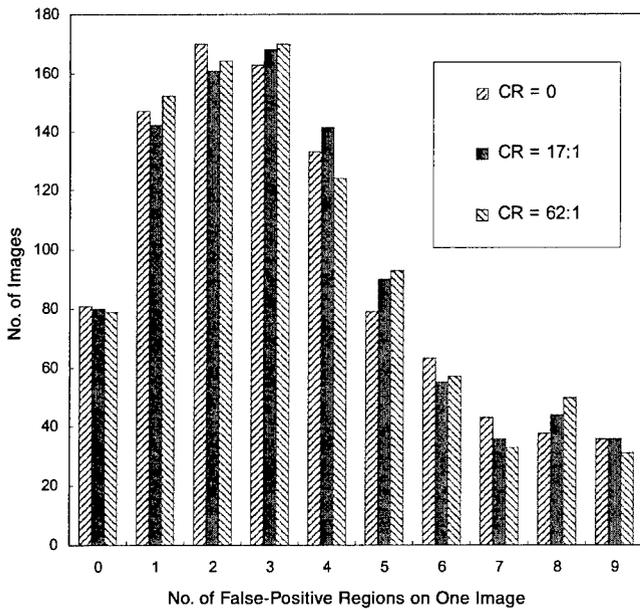
Note.—Regions with suspected microcalcification clusters identified with CAD before application of the ANN (20).

false-positive detections. The variance term in the one-way analysis of variance was adjusted for correlations resulting from the fact that images were evaluated in all six modes. For a comparison of the true-positive detections across the six modes, we applied the same test for trend to the proportion of images for which all positive regions had been correctly identified. For each mode and for both mass and microcalcification clusters, the  $\chi^2$  goodness-of-fit test was used to determine whether the distribution of false-positive detections per image obeyed a Poisson distribution.

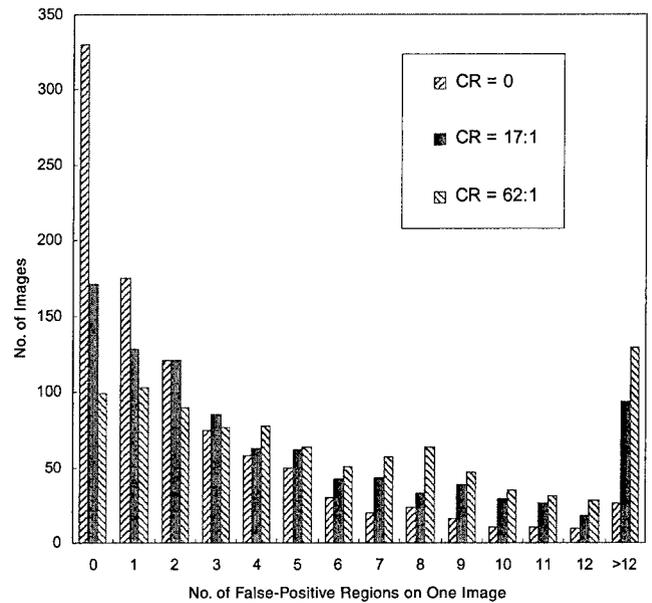
After application of the ANN, a score from 0 to 1 is assigned to each identified mass or microcalcification cluster. Equality of the average ANN score per image across the six modes was compared separately for the true-positive regions and the false-positive regions by using a two-way analysis of variance (mode or compression level by image). When the scores for true-positive regions were compared across modes, if a true-positive region was not identified, it was assigned a score of 0.

## RESULTS

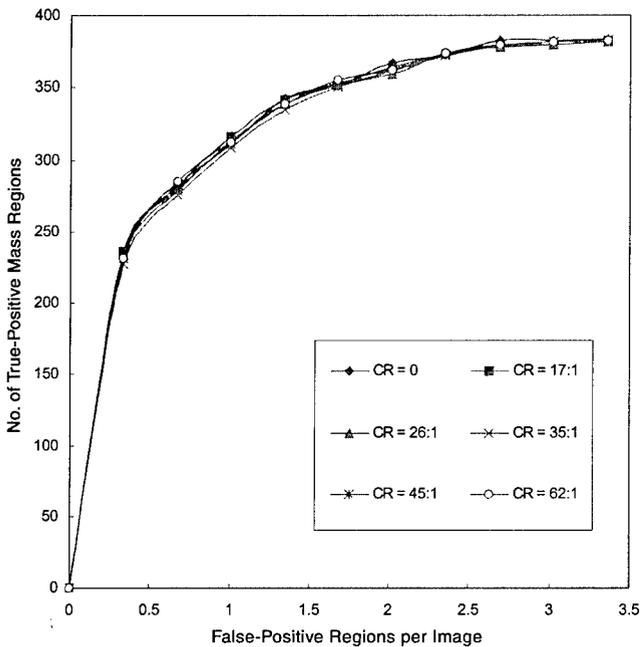
Tables 1 and 2 summarize the performance of the CAD schemes in identifying regions suspicious for masses or microcalcification clusters before the application of ANNs. Results are summarized for noncompressed images and the five compression modes and include the total number of true-positive regions detected, the false-positive detection rates, and the fraction of images without any false-positive detections. Figures 2 and 3 show the distribution of false-positive regions per image before image compression and when images had been compressed to five levels (average compression ratios, 17:1 and 62:1). Table 1 and Figure 2 demonstrate that JPEG image compression had little impact on mass detection before the ANN was applied. The maximum change in the number of true-positive regions identified for the six modes is less than 0.5% (between 381 and 383 regions for all modes,  $P = .31$ ). There was no significant difference in either the number of images without a



**Figure 2.** The number of false-positive mass regions identified in three testing modes (at no compression and at compression ratios [CR] of 17:1 and 62:1), before application of the ANN.



**Figure 3.** The number of false-positive microcalcification cluster regions identified before the ANN was applied in three testing modes (at no compression and at compression ratios [CR] of 17:1 and 62:1).



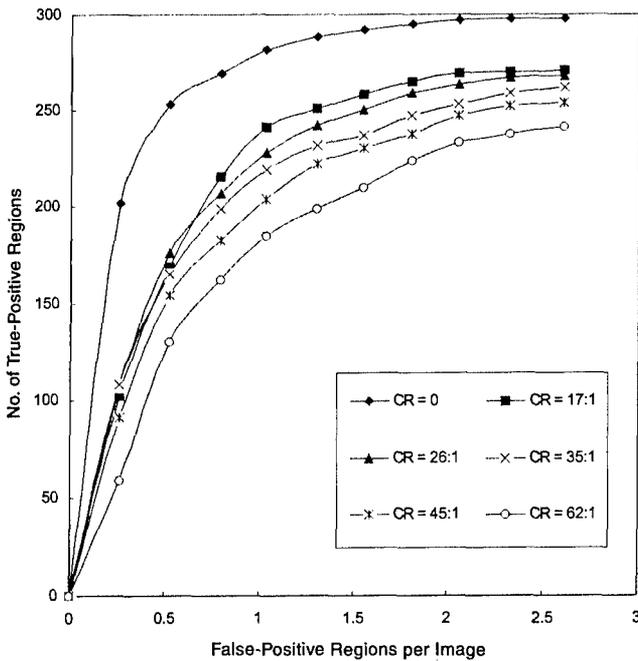
**Figure 4.** The number of detected true-positive mass regions at six image-compression levels as a function of the false-positive rates. CR = compression ratio.

false-positive detection ( $P = .50$ ) or the average number of false-positive regions detected per image ( $P = .91$ ).

The image-compression schemes did affect the detection of microcalcification clusters. Table 2 and Figure 3 demon-

strate a reduction in initially identified true-positive cluster regions from 298 for noncompressed images to as few as 269 (9.7% reduction) for 45:1 compression ( $P < .001$ ). As the compression levels increased there was a corresponding decrease in the proportion of images without any false-positive regions ( $P < .001$ ). The initially identified false-positive rate of 2.61 regions per image for the non-compression mode increased to 6.27 (140% increase) for the 62:1 compression ratio ( $P = .13$ ). When the distributions of false-positive regions per image (Figs 2, 3) were tested to determine whether the data were in a Poisson distribution, the hypothesis of such a distribution was rejected for both mass detection and clusters of microcalcification detection for all modes. This occurred because the number of images with very few (two or fewer) or many (seven or more) false-positive regions was much higher than expected under the assumption of a Poisson distribution.

When the average scores assigned by the ANN were compared across modes for the true-positive regions, there was a clear decrease in scores for microcalcification cluster regions as the compression increased ( $P < .001$ ), but for masses there was no significant difference ( $P = .25$ ). A higher average ANN score is associated with a higher true-positive fraction for a given false-positive fraction. Figures 4 and 5 show the number of true-positive regions detected at selected levels at the average number of false-positive detections per image (0.25 or more). When the average scores assigned by the ANN for

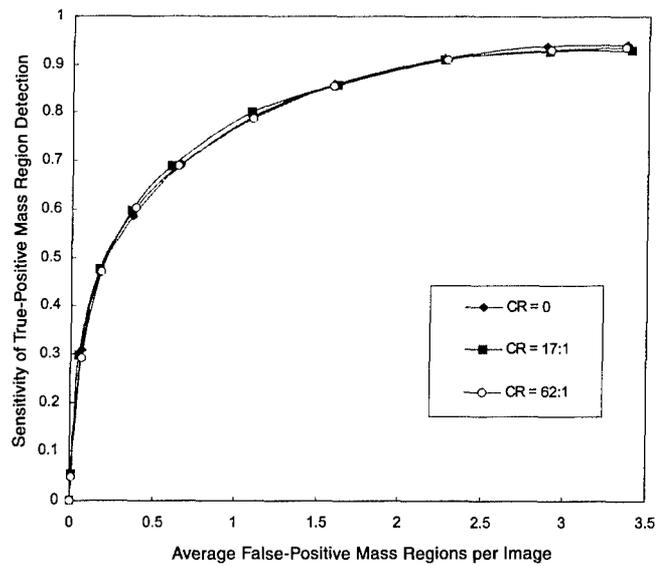


**Figure 5.** The number of detected true-positive microcalcification clusters at six image-compression levels as a function of the false-positive rates. CR = compression ratio.

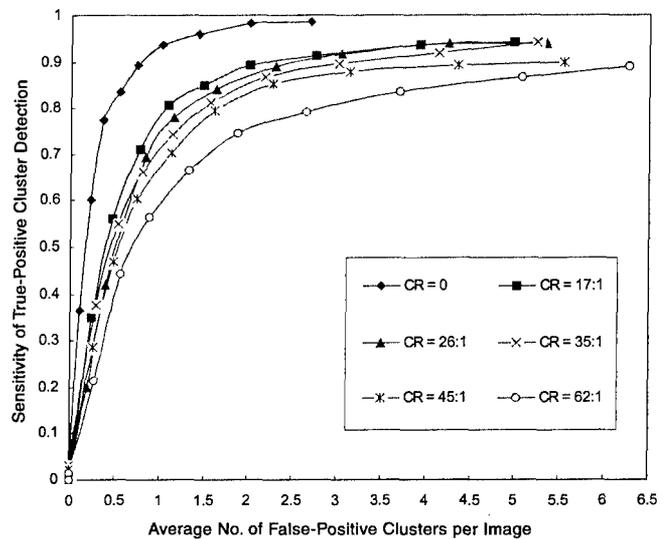
false-positive regions were compared across modes, there was no difference for masses ( $P = .40$ ), but there was a significant increase in scores for regions suspected to have microcalcification clusters as the compression levels increased ( $P < .001$ ). The results were similar when we analyzed change in patterns of the false-positive detection rates at a series of fixed sensitivity levels for all compression modes.

Figures 6 and 7 show a series of free-response receiver operating characteristic (FROC) curves. These curves demonstrate the ultimate performance (after application of ANN) when our CAD schemes were used to detect mass and microcalcification cluster regions in the testing database containing 952 images under six different compression modes. The CAD performance for the detection of masses shows little change among the six testing modes (Table 1, Fig 4), so only three FROC curves were plotted in Figure 6. Because 408 true-positive mass regions were included in the testing database and the CAD scheme identified only 383 (Table 1), the maximum detection sensitivity is 94%, as shown in Figure 6. From the FROC curves we also find that our CAD scheme can identify 80% of true-positive mass regions at an average false-positive detection rate of about one region per image for both noncompressed images and all five JPEG compression modes.

As shown in Figure 7, the maximum sensitivity for microcalcification cluster detection can reach 98%. For a detection



**Figure 6.** Performance of the CAD scheme, including the ANN, for mass detection ( $n = 952$  images). CR = compression ratio.



**Figure 7.** Performance of the CAD scheme, including the ANN, for microcalcification cluster detection ( $n = 952$  images). CR = compression ratio.

sensitivity of 80%, the false-positive detection rate is 0.4 for noncompressed images and increases to 1, 1.2, 1.4, 1.5, and 3 regions per image for the five JPEG compression levels from 17:1 to 62:1. Figure 7 also demonstrates that some FROC curves may cross (eg, curves for compression ratios of 26:1 and 35:1) at low sensitivity levels (<60%). For sensitivity levels above 70%, which have more practical value, no curves cross. Figure 7 shows that as the compression ratio increases, there is a monotonic trend for simultaneous decreases in detection sensitivities and increases in false-positive detection rates.

## DISCUSSION

In this study we assess the relationship between JPEG-type image compression and performance changes in our CAD scheme for both mass and microcalcification cluster detection. The JPEG, a lossy data-compression algorithm, is based on the concept of gradually compromising the accuracy of the reconstructed image in exchange for increased data compression (29). The unique characteristics of this study are that (a) we used a wide range of compression ratios, from 17:1 to 62:1, and (b) we tested the impact of these compression levels on CAD results in a large database with 952 images.

It should be noted that for mass detection the images are first subsampled (pixel averaged) by a factor of four in two dimensions, which is approximately equivalent to the 16:1 data compression. The effect generated by pixel averaging, however, is different from that generated by JPEG data compression, for both frequency-dependent information content and noise characteristics in the decompressed (reconstructed) image. Our previous "just-noticeable difference," or JND, study (using the same quantization factors) indicated that the JPEG scheme imposed barely noticeable, albeit physically measurable, distortion for most reconstructed images (8). Such distortions could be transferred to the subsampled images, but these changes (both noise and distortion) have been found to have virtually no impact on the performance of the CAD scheme for mass detection. Since compression-generated changes are not generally in the frequency range of interest for mass detection, we did not observe noticeable changes in the number of initially selected suspicious regions at all the compression levels. The topographic multilayer region growth (23,26) is performed on the reconstructed images. Hence, the feature values computed for the regions and the background may be different in each testing mode. Any difference in the feature vector is transferred into the input neurons of the ANN. In this study, we found that such differences did not substantially affect the classification output of the ANN. This indicates that for mass detection our training protocol (25) for the ANN did not result in overtraining and was quite robust to the JPEG compression-generated noise.

Although results of another study indicated that data compression should be limited to less than 9.6:1 (6) for the detection of a single microcalcification, our results are not comparable. First, we used different image databases and compression methods. Most important, we assessed the detection of different targets, namely, microcalcification clusters versus single microcalcifications. Our results showed that both sensitivity and specificity were affected in the detection of microcalcification clusters. While the sensitivity

results are somewhat encouraging (<5% decrease up to 35:1 compression ratios), the specificity decreases markedly even at low compression ratios. These results indicate that even at high levels of compression, the JPEG algorithm can preserve some information on the presence of microcalcification clusters in the decompressed images. The marked increase in identified false-positive clusters could mean that our previous CAD scheme, as optimized, might be too customized to a specific set of image (or data) characteristics. Not surprisingly, images with more false-positive clusters identified before compression ("difficult" images) seem to degrade faster with increasing levels of compression, with the number of identified clusters increasing rapidly.

We emphasize that both CAD schemes used in this study had been trained and optimized with a set of noncompressed images. If these or similar schemes are implemented with picture archiving and communication systems or telemammography applications, they should ideally be applied to full-fidelity images than to the compressed data sets. Furthermore, compression-specific optimization is required before these schemes can reliably be applied to highly compressed images.

Finally, because CAD performance may depend on the difficulty and diversity of the training database (18), as well as the potential bias in the feature domain and possibly overfitting during training (24), this study serves indirectly as a robustness test for the independent noncompressed data set. In this regard, the experimental results with the noncompressed images were encouraging.

## ACKNOWLEDGMENTS

The authors thank William Reinus, MD, at Washington University, St Louis, Mo, and the medical imaging research and development group at Eastman Kodak Company, Rochester, NY, for providing us with images used in this study.

## REFERENCES

1. Strax P. Detection of breast cancer. *Cancer* 1990; 66:1336-1340.
2. Miller AB. Mammography: reviewing the evidence-epidemiology aspects. *Can Fam Physician* 1993; 39:85-90.
3. Feig SA, D'Orsi CJ, Hendrick RE. American College of Radiology guidelines for breast cancer screening. *AJR* 1998; 171:29-33.
4. Lou SL, Sickles EA, Huang HK, et al. Full-field direct digital telemammography: technical components, study protocols, and preliminary results. *IEEE Trans Inf Technol Biomed* 1997; 1:270-277.
5. Maitz GS, Good WF, Gur D, et al. Preliminary clinical evaluation of a high-resolution telemammography system. *Invest Radiol* 1997; 32:236-240.
6. Chan HP, Lo SB, Niklason LT, Ikeda DM, Lam KL. Image compression in digital mammography: effects on computerized detection of subtle microcalcifications. *Med Phys* 1996; 23:1325-1336.
7. Goldberg MA, Pivovarov M, Mayo-Smith WW. Application of wavelet compression to digitized radiographs. *AJR* 1994; 163:463-468.
8. Good W, Maitz GS, Gur D. Joint Photographic Experts Group (JPEG) compatible data compression of mammograms. *J Digit Imaging* 1994; 7:123-132.

9. Vyborny CJ, Giger ML. Computer vision and artificial intelligence in mammography. *AJR* 1994; 162:699-708.
10. Nishikawa RM, Giger ML, Schmidt RA, Wolverton MD, Collins SA, Doi K. Computer-aided diagnosis in screening mammography: detection of missed cancers (abstr). *Radiology* 1998; 209(P):335.
11. Nawano S, Murakami K, Moriyama N, Kobatake H. Computer-aided diagnosis in full digital mammography. *Invest Radiol* 1999; 34:310-316.
12. Vyborny CJ. Can computers help radiologists read mammograms? *Radiology* 1994; 191:315-317.
13. Hoffman KR. For the proposition in the point/counterpoint: 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.
14. Mendez AJ, Tahocas PG, Loda MJ. Computer-aided diagnosis: automatic detection of malignant masses in digitized mammograms. *Med Phys* 1998; 25:957-964.
15. Chan HP, Sahiner B. Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. *Med Phys* 1998; 25:2007-2019.
16. Huo Z, Giger ML, Vyborny CJ, Doi K. Automated computerized classification of malignant and benign masses on digital mammograms. *Acad Radiol* 1998; 5:155-168.
17. Zheng B, Chang YH, Gur D. On the reporting of mass contrast in CAD research. *Med Phys* 1996; 23:2007-2009.
18. Nishikawa RM, Giger ML, Doi K. Effect of case selection on the performance of computer-aided detection schemes. *Med Phys* 1994; 21:265-269.
19. Wei D, Chan HP, Helvie MA, et al. Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis. *Med Phys* 1995; 22:1501-1513.
20. Kundel HL, Revesz G. Lesion conspicuity, structure noise, and film reader error. *AJR* 1977; 126:1233-1238.
21. Zheng B, Chang YH, Good WF, Gur D. Assessment of mass detection using tissue background information as input to a computer-assisted diagnosis scheme. *Proc SPIE* 1998; 3338:1547-1555.
22. Revesz G, Kundel HL, Toto LC. Densitometric measurements of lung nodules on chest radiographs. *Invest Radiol* 1981; 16:201-205.
23. 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.
24. 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.
25. Zheng B, Chang Y-H, Wang X-H, Good WF. Applying a genetic algorithm for the improvement of decision making in medical image diagnosis. Presented at the IASTED International Conference on Artificial Intelligence and Soft Computing, Honolulu, Hawaii, USA, August 11, 1999.
26. 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.
27. Chang YH, Zheng B, Gur D. Computer-aided detection of clustered microcalcifications on digitized mammograms: a robustness experiment. *Acad Radiol* 1997; 4:415-418.
28. Abelson RP, Tukey JW. Efficient utilization of non-numerical information in quantitative analysis: general theory and the case of simple order. *Ann Math Stat* 1963; 34:1347-1369.
29. Gonzalez RC, Woods RE. *Digital image processing*. Reading, Mass: Addison-Wesley, 1992.

**Soft-Display Mammographic Readings Under Different  
Computer-Assisted Detection Cueing Environments: Preliminary Findings**

Bin Zheng, Ph.D.

Marie A. Ganott, M.D.

Cynthia A. Britton, M.D.

Christiane M. Hakim, M.D.

Lara A. Hardesty, M.D.

Thomas S. Chang, M.D.

Howard E. Rockette, Ph.D.

David Gur, Sc.D.

Department of Radiology, University of Pittsburgh,  
Pittsburgh, PA 15261-0001 and  
Magee-Womens Hospital, University of Pittsburgh Medical Center Health System,  
Pittsburgh, PA 15213

This work is supported in part by the U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD, 21702-5014 under Contracts DAMD17-98-1-8018 and DAMD17-00-1-0410. 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. This work is also supported by grant CA77850 from the National Cancer Institute, National Institutes of Health.

Corresponding Author:  
Reprint Address:

Bin Zheng, Ph.D.  
Imaging Research, Suite 4200  
Magee Womens Hospital  
300 Halket Street  
Pittsburgh, PA 15213  
Phone: 412/641-2568  
Fax: 412/641-2582  
Email: [bzheng@radserv.arad.upmc.edu](mailto:bzheng@radserv.arad.upmc.edu)

**Original Research**

**Soft-Display Mammographic Readings Under Different  
Computer-Assisted Detection Cueing Environments: Preliminary Findings**

## ABSTRACT

**Purpose:** To assess the performance of radiologists when detecting masses and microcalcification clusters on digitized mammograms using different Computer-Assisted Detection (CAD) cueing environments.

**Materials and Methods:** 209 digitized mammograms depicting a total of 57 verified masses and 38 microcalcification clusters in 85 positive and 35 negative cases were interpreted independently by seven radiologists using five different display modes. Except the first mode, for which no CAD results were provided, suspicious regions identified by a CAD scheme were cued in all other modes using a combination of two cueing sensitivities (90% and 50%) and two false-positive rates (0.5 and 2 per image). A receiver-operating characteristic (ROC-type) study was carried out using soft display.

**Results:** CAD cueing at 90% sensitivity and 0.5 false-positive regions per image improved observers' performance levels significantly. As accuracy of CAD cueing decreased so did observer performances ( $P < 0.01$ ). Cueing specificity affected mass detection more significantly, while cueing sensitivity affected the detection of microcalcification clusters more significantly ( $P < 0.01$ ). Reducing cueing sensitivity and specificity significantly increase false-negative rates in non-cued areas ( $P < 0.05$ ). Trend results were consistent for all observers.

**Conclusion:** CAD systems have the potential of significantly improving diagnostic performance in mammography. However, poorly performing schemes could adversely affect observer performance in both cued and non-cued areas.

**Key Words:** Breast Cancer, Observer performance study, Computer-assisted detection, Mammography.

## INTRODUCTION

Breast cancer is one of the leading causes of death in women over the age of 40 [1,2]. To reduce mortality and morbidity of patients through early diagnosis and treatment, current guidelines recommend periodic mammography screening for women age forty and over [3]. Due to the large volume of mammograms performed and the low yield of abnormalities in screening environments, detecting abnormalities (mainly masses and microcalcification clusters) from the background of 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 a valuable "second opinion" to radiologists and aiding could help improve the 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 using CAD-prompted systems. These studies 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 benign ones [10-12]. While some studies indicated that using CAD did not significantly decrease the specificity levels of the radiologists [13-15], others indicated that current CAD systems could significantly decrease radiologists' diagnostic accuracy and efficiency due to the high false-positive detection rates [16,17]. Similar to the difficulty in comparing the performance of different CAD schemes developed at various institutions [18], the results of these studies are not easily compared since different CAD schemes, radiologists, and cases were included. These studies did not address in

detail how CAD performance could affect observers' diagnostic performance or the level of CAD performance that may be required in order to be widely acceptable as a true aiding tool in the clinical environment. Researchers have suggested that large-scale experiments are needed to assess the effect of CAD performance (e.g., the false-positive identifications) on the diagnostic accuracy of radiologists [19]. Some doubt remains whether using CAD systems might increase the number of unnecessary follow-up examinations or biopsies, thereby offsetting the benefits from the potential gains in sensitivity [20].

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

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

## MATERIALS AND METHODS

Seven board-certified radiologists with a minimum of three years' experience in the interpretation of mammograms participated in this observer performance study. None of these 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 in our laboratory under an IRB-approved, patient-consent exempt protocol. The original database contained mammograms collected mainly from several thousand patients undergoing routine mammographic screening in three different medical centers [27]. All positive masses were biopsy verified. All the negative cases were rated as to level of concern by radiologists using standard BI-RADS recommendations. The negative cases had been diagnosed as negative during at least two subsequent follow-up examinations. Although we routinely acquire four images in a single examination (2 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. Using an established digitization protocol, all mammograms were digitized using a laser-film digitizer (Lumisys, Sunnyvale, CA) with a pixel size of  $100\ \mu\text{m} \times 100\ \mu\text{m}$  and 12-bit digital-value resolution. The quality of the digitizer was monitored routinely to ensure that value levels were linearly proportional to optical density in the range of 0.2 to 3.2 [28].

The selection of "subtle" or "difficult" cases includes several steps. First, we select a large set of positive cases (in this experiment 200) for which the output scores generated by the CAD scheme are low for the likelihood that the abnormality in question is present [27]. Similarly, a set of suspicious negative cases (in this experiment 80) is used for which CAD scores were high for the likelihood that a mass or a cluster of microcalcifications, or both was present. Then, two experienced observers prune the data set by visual inspection on the same display as used in the study with the "true diagnosis" known to select the final 120 cases to be used in the study. The total number of positive cases was

selected to include a reasonable mix of benign and malignant cases depicting both single and multiple abnormalities with a minimum of 25 malignant cases depicting each of the abnormalities. The resources required in terms of radiologist effort (reading time) was a factor in limiting the total number of cases in this study to 120 and reading modes to 5. Of these, 85 depicted either masses or clusters of microcalcifications, or both, and 35 cases were negative for these abnormalities. Ten of the positive cases depicted both a mass and a microcalcification cluster. All other positive cases depicted only one abnormality (either a mass or a cluster). Hence, the positive cases consisted of a total of 38 verified microcalcification clusters and 57 verified masses. Biopsy results indicated that 27 of the clusters and 39 of the 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 based on subtleness of the depicted abnormality, and no attempt was made to balance the number of benign and malignant cases in the dataset. Although studies suggested that in order to preserve subtle microcalcifications mammograms should be digitized using pixel sizes of  $50 \mu m \times 50 \mu m$  or less [15,29], all the microcalcification clusters in this study were detectable by our CAD scheme. In addition, we verified that all these clusters were visible on the images when digitized with  $100 \mu m \times 100 \mu m$  pixel size.

In this study, radiologists were asked to detect masses and microcalcification clusters in digitized mammograms displayed on a monitor. In most of the 120 cases (89), two contralateral images (the same view of left and right breasts) were displayed on the monitor side-by-side. For some cases (31), only a single image was displayed. The latter group was selected from the cases for which we have only two views of one breast in our database. Hence, only one view was displayed in this study following our study protocol. Table 1 summarizes the distribution of the abnormalities depicted in these 120 cases by type and verified finding. 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 were made available during the interpretation.

Each radiologist interpreted the same 120 cases five times using five different display modes. With the exception of the first mode in which no CAD results were provided to the radiologists, suspicious regions, as identified by our CAD schemes, were marked (cued) on the images in all other modes. Two true-positive cueing sensitivity levels (90% and 50%) and two false-positive cueing rates (0.5 or 2 per image) were used in these four cueing modes (see Table 2). During the cued modes, when a new case was loaded onto 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 schemes use filtering, subtraction, and topographic region growth algorithms to identify suspicious regions (including masses and microcalcification clusters) [30,31]. Then, using nonlinear multi-layer multi-feature analyses, two pre-trained artificial neural networks (ANNs) were used to classify each region as positive or negative for the presence of an abnormality in question [32]. One was designed to assess regions suspicious for masses and the other one was for microcalcification clusters. Before applying the ANNs, 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 [95/209] 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 774/209). The ANNs were then applied to classify all of these regions. Each suspicious region received a likelihood score for being positive (from 0 to 1). 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 cueing sensitivities (90% and 50%) were applied to masses and microcalcification clusters. Each abnormality was assigned a number (e.g., from 1 to 57 for masses or 1 to 38 for clusters). A computer program randomly selected regions to be cued until the required number was reached for the sensitivity level being evaluated. In display modes #2 and #3 with the cueing sensitivity set at 90%, 51 true masses of 57 and 34 of 38 clusters were selected. In modes #4 and #5 with the cueing sensitivity set at 50%, 29 of the 57 masses and 19 of the 38 clusters were selected. Two false-positive cueing rates (approximately 0.5 and 2 false-positive regions per image) were used. Because the total number of false-positive "clusters" identified by the scheme was 95, all of these regions were used in display modes #3 and #5, which provided a false-positive cueing rate of 0.45 (95/209). In modes #2 and #4, the total false-positive desired cueing rate was 0.5 per image, which was one fourth of that in modes #3 and #5. Hence, one-fourth (24) of the available (95) false-positive "clusters" were selected based on the ANN-generated scores with the 24 highest scoring regions being selected in descending order, resulting in a cueing rate of 0.11 (24/209). To reach the overall target of 0.5 and 2 false-positive cues per image (including both mass and microcalcification cluster regions), 774 false-positive mass regions were also sorted based on the ANN-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 modes #3 and #5. Thus, the false-positive cueing rates for mass only were 0.39 (82/209) and 1.55 (324/209) per image, respectively. In summary, modes #2 and #4 included 106 (24+82) false-positive cues (or 0.5 per image), and modes #3 and #5 included 419 (95+324) false-positive cues (or 2 per image).

Each of the 20 reading sessions for individual observers included 30 randomly selected cases using one reading mode. To eliminate the potential for learning effects, the order of display modes (or cueing rates) for each observer was pre-selected using a counterbalanced approach. The 20 sessions were divided into 4 blocks with 5 sessions each. In each block, one observer read five sessions with five different modes in a random order. However, at each session number in the series (e.g., session #6), at least five observers read different modes, and no more than two readers read 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 to 5 read modes 1 to 5, respectively, while observer 6 read mode #3 and observer 7 read 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 (e.g., session #4) and their sequential order for each observer were different. A minimum time delay (10 days) between two consecutive readings of the same case was implemented.

A standard SUN SPARC-20 landscape workstation monitor was used to display the images. Images were not pre-processed other than we did optimize the contrast of each individual image through a window and level manipulation for optimal visual display. The image parameters were then fixed. The observers could not manipulate the contrast and brightness during the readings. Initially, images were displayed on the screen as sub-sampled (low resolution) to fit the screen size (with approximately  $1,200 \times 850$  pixels). Using zoom and roam functions, the radiologists were able to view the images at full 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 sub-sampled images or full-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 they were required to classify the suspected abnormality as benign or malignant. Once a radiologist pointed to and clicked the cursor onto the center of a suspected abnormality, a scoring window appeared, followed by a confidence level sliding scale. The program automatically recorded all diagnostic information entered by the radiologist, including the type of a detected abnormality (mass or microcalcification cluster), location (the center of the detected region), and two estimated likelihood scores (from 0 to 1) for detection (presence/absence) and for classification (benign/malignant) of any identified region that was suspected for depicting an abnormality. The likelihood scores were used to generate FROC curves.

The results for each observer, each abnormality, and each display mode were qualitatively viewed, and FROC 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 FROC curves (or the detection sensitivities at the same false-positive rates) across four different CAD cueing modes, we compared sensitivities among curves at ten different false-positive rates uniformly distributed over the measured range. Sensitivity levels across modalities were compared 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 using a Generalized Estimating Equation (GEE) approach [33]. In addition, we analyzed the changes of performance indices (i.e., the number of missed true-positive regions in the cued or non-cued areas) for the two sensitivity levels (50% and 90%) and for the two false-positive cueing rates (0.5 to 2 per image). The hypotheses of equality of the number of missed abnormalities were also tested using a repeated measures logistic regression with reader and modality in the model. Last, to examine the potential biases for reading the same case five times, the reading

results were re-ordered and analyzed for all cases read the first time (regardless of mode) as one group, and all cases read for the second time as another group, etc. Performance curves were computed separately for these five mutually exclusive groups and were compared (using the analysis of variance test).

## RESULTS

Performance curves varied among observers, but the general pattern was consistent for all observers. Figures 1 to 3 demonstrate the average performance of the seven observers. These figures present curves of the average performance for the detection of either abnormality, masses alone, or microcalcification clusters alone, respectively. As noted from the non-cued results (mode #1), the task in general was challenging, whether due to 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 between modes #2 through #5 were highly significant ( $P < 0.01$ ). However, the results showed different patterns for the detection of masses as compared with microcalcifications. In the case of masses (Figure 2), specificity of the CAD results (or cueing false-positive rate) affected the observer in a more significant manner. The differences between modalities was statistically significant ( $P < 0.01$ ) with the performance decreasing as the total number of cued regions increases. In the case of clusters (Figure 3), observers' performances were affected to a greater extent by the cueing sensitivity. The combination of case subtlety and viewing on soft display rendered the test of microcalcification cluster detection so difficult that only approximately 60% were

detected without cueing or with cueing at low sensitivity (modes #4 and #5). With the support of highly sensitive cues, the performance improved to a detection rate of approximately 75% ( $P < 0.01$ ).

Highly accurate cueing (i.e., 90% sensitivity and 0.5 false-positive cues per image) helped the observers improve performance as compared with the non-cued environment ( $P < 0.01$ ). As the accuracy of the cueing decreases, so does the performance of the typical observer. This effect continues for either detection task, but the detection of microcalcification clusters was more significantly affected by sensitivity of the cueing in our case. Most important, perhaps, our results clearly indicate that overall poorly performing CAD (Figure 1) can result in significant degradation of observer performance ( $P < 0.01$ ).

Table 3 demonstrates the number of CAD-cued abnormalities that were identified in mode #1 (non-cueing) but were missed in other (cued) modes by each radiologist. Some increases in rejection rates of true-positive regions were observed when the total number of cues increased, but the results were not significant ( $P > 0.05$ ).

Table 4 summarizes the number of missed abnormalities in non-cued areas during CAD-cued observations. The table shows that for the highly sensitive cueing modes (e.g., modes #2 and #3, where only 10% of true-positive regions were not cued), the majority of the missed abnormalities (> 94%) were also missed in mode #1. As CAD cueing sensitivity is reduced to 50%, the average number of missed abnormalities in non-cued areas increased significantly ( $P < 0.05$ ). More importantly, approximately 30% of these regions were detected by the radiologists in mode #1. Increasing false-positive cueing rate from 0.5 to 2 per image (mode #4 vs mode #5) increased the number of missed abnormalities in non-cued areas from an average of 14.4 to 18.0, which was not significant ( $P = 0.16$ ), most likely due to the small sample size. In this case, the observers also missed significantly more

regions that were detected in mode #1 ( $P=0.03$ ). In general, the number of missed abnormalities (false-negative rate) in the non-cued areas increases as the cueing sensitivity decreases and false-positive cueing rate increases. As a result, mode #5 has the highest miss rate in non-cued areas. When we compared the 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 that 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 ROC curves ( $A_z$ ) for the different modes (ROCFIT [34]). Areas under the curves were estimated using maximum likelihood (MLE) under the binormal assumption. Areas under the ROC curves for classification performance over all readers were  $0.70\pm 0.02$ ,  $0.69\pm 0.02$ ,  $0.69\pm 0.02$ ,  $0.70\pm 0.02$ , and  $0.68\pm 0.02$  for modes #1 through #5, respectively. Comparing each pair of modes did not result in any significant differences ( $P>0.05$ ). Hence, once identified (detected), the observers' ability to distinguish between benign vs malignant abnormalities (classification) were not significantly affected ( $P>0.05$ ) by the cueing mode or lack thereof. Although there were differences in performance among the observers, we did not identify any correlation for either the detection or classification tasks with observers' 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 days to 154 days, with an average time delay of 48 days. When we examined the results after re-ordering cases by their order of

appearance (i.e., first time, second time), regardless of the mode, no significant difference between the groups ( $P>0.8$ ) was identified (Figure 4). Similar performance patterns were observed when the 31 cases that included only one image were excluded from the analyses, and the detection results were not significantly altered in any comparison between the results for the whole group (120 cases) and the subset of 89 cases containing two images ( $p>0.5$ ).

## DISCUSSION

This preliminary study under laboratory conditions has to be clearly viewed as such. The fact that the conditions in the study were removed from the typical clinical environment has to be considered before any generalization of the results is contemplated. However, the consistency of the patterns observed for the individual readers and the group as a whole warrant further assessments 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 in question 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 non-cueing performance curve. This observation is consistent with some of the results previously reported by others, but it may not be clinically relevant in situations when most abnormalities are not as difficult to detect as those in this study.

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 routine clinical practice. In particular, one should consider the cueing performance level of the scheme itself and the potential increase in missed abnormalities in non-cued regions due to the fact that 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 cueing 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 using current levels of CAD performances as well as what one hopes to achieve using CAD in the future. The range of CAD performances used for cueing 90% sensitivity at 0.5 false-positive identifications per image to 50% sensitivity at 2 false-positive identifications per image clearly make this study an interesting one in enabling an assessment of what could be expected under improved CAD results. It is interesting to note that for all display modes, the use of CAD cueing with either high or low performance had a limited effect on observers when they operated on a conservative level. Namely, they indicated only regions they were quite confident about and therefore had low false-positive rates. This stemmed largely from the fact that the CAD cueing identified mainly truly appropriate ("reasonable") areas on the image as "suspicious." As observers loosened their criteria (indicated a larger number of suspicious regions), the CAD-cueing performance affected observers in a more significant manner. Namely, the use of the better performing cueing scheme significantly improved observer performance, while the use of the poorly performing cueing schemes significantly degraded observer performance.

Analysis of the datasets after reordering cases by appearance indicate that “learning” effects, if any, were not a significant factor in this study. Although all selected abnormalities in this study were detectable by the CAD schemes and visible on the 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 using soft display for this purpose. We are not aware of any comprehensive study assessing this issue, and our results, albeit very preliminary, suggest that such a study should be performed.

Despite the limited information provided (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 cueing performance, pointing 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 have been shown to perform well [12] 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 to 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-display using a limited number of views without prior examinations being available for comparison. We note the difference between this and other reported studies where observers could view both hard copy images and low-resolution soft copy images with CAD-cued areas on the screen [14,15]. Not providing hard copy images to the observers

could be a significant factor in lowering detection sensitivity in this study. This resulted in a crossing of the performance curves for the detection of microcalcifications (Figure 3), since the non-cued mode exhibited a “capping” effect (an imposed upper limit) that was “removed” with the aid of CAD cueing. This does not invalidate any of the analyses or observations made in this study. Despite the generally low level of performance and the fact that we used very high prevalence of abnormalities in our dataset, 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) each session. The effects we observed under these conditions are probably different and likely minimized as compared with a study design in which each mode is read to its completion before any prevalence changes (i.e., change to a different mode).

In conclusion, our preliminary study indicates that in a laboratory environment, observer performance in the detection of subtle mammographic abnormalities is significantly affected by the inherent performance of a cueing system. High performance cueing systems can significantly improve observer performance. On the other hand, low performance cueing systems can significantly degrade observer performance. These findings, together with the inter-mode 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 datasets 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* 1998; 171:29-33.
4. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. *Radiology* 1992; 184:613-617.
5. Thurffjell 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* 1994; 162:699-708.
7. Hoffman KR. For 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; 26:1-2.
8. Nishikawa RM, Giger ML, Schmidt RA, Wolverton DE, Collins SA, Doi K. Computer-aided diagnosis in screening mammography: detection of missed cancers. *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, ed. *Computer-aided diagnosis*. Elsevier Science B.V., **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 (Article in German, English abstract can be founded in MEDLINE <http://www.nlm.nih.gov/databases/freemedl.html>).
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 (Article in German, English abstract can be founded in MEDLINE <http://www.nlm.nih.gov/databases/freemedl.html>).
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. *Human Factors* **1984**; 26:223-234.
22. Krose BA, Julesz B. The control and speed of shifts of attention. *Vision Research* **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. *Behavior Research Methods, Instruments & Computers* **1992**; 24:475-485.
26. Krupinski EA, Nodine CF, Kundel HL. Perceptual enhancement of tumor targets in chest x-ray images. *Perception & Psychophysics* **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 multiplayer 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. *Stats in Med* **1998**; 17:1033-1053.

## List of Table Captions

Table 1: Number of mammographic cases in different categories. (M – malignant, B – benign).

Table 2: CAD cueing conditions of the five display modes used in the study.

Table 3: The number of missed abnormalities that were identified as suspicious in mode 1 (non-cued) but missed in other modes despite the fact that the abnormality in question was cued.

Table 4: The number of missed abnormalities in non-cued regions. The number in parenthesis indicates the number of missed regions that were detected in mode 1 (non-cued).

## List of Figure Captions

Figure 1: Curves of average detection performance of mammographic abnormalities (including both masses and microcalcification clusters) for seven participating radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (\*), and mode 5 (◆).

Figure 2: Curves of average performance of mass detection for seven radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (\*), and mode 5 (◆).

Figure 3: Curves of average performance of microcalcification cluster detection for seven radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (\*), and mode 5 (◆).

Figure 4: Curves of average detection performance of abnormalities for seven radiologists as a function of the order of appearance or round (e.g., first time, second time, etc) and regardless of reading mode. Order of appearance is represented as follows: first time (o), second time (■), third time (▲), fourth time (\*), and fifth time (◆).

Table 1: Number of mammographic cases in different categories. (M – malignant, B – benign).

	Mass		Microcalcification cluster		Both mass and cluster		Negative	Total cases
	M	B	M	B	M	B		
Single image cases	10	1	11	3	1	1	4	31
Two image cases	20	16	7	7	8	0	31	89
Total Cases	30	17	18	10	9	1	35	120

Table 2: CAD cueing conditions of the five display modes used in the study.

Reading mode	CAD cueing	Cueing sensitivity	Cueing FP rate
1	No		
2	Yes	0.9	0.5
3	Yes	0.9	2
4	Yes	0.5	0.5
5	Yes	0.5	2

Table 3: The number of missed abnormalities that were identified as suspicious in mode 1 (non-cued) 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: The number of missed abnormalities in non-cued regions. The number in parenthesis indicates the number of missed regions that were detected in mode 1 (non-cued).

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)

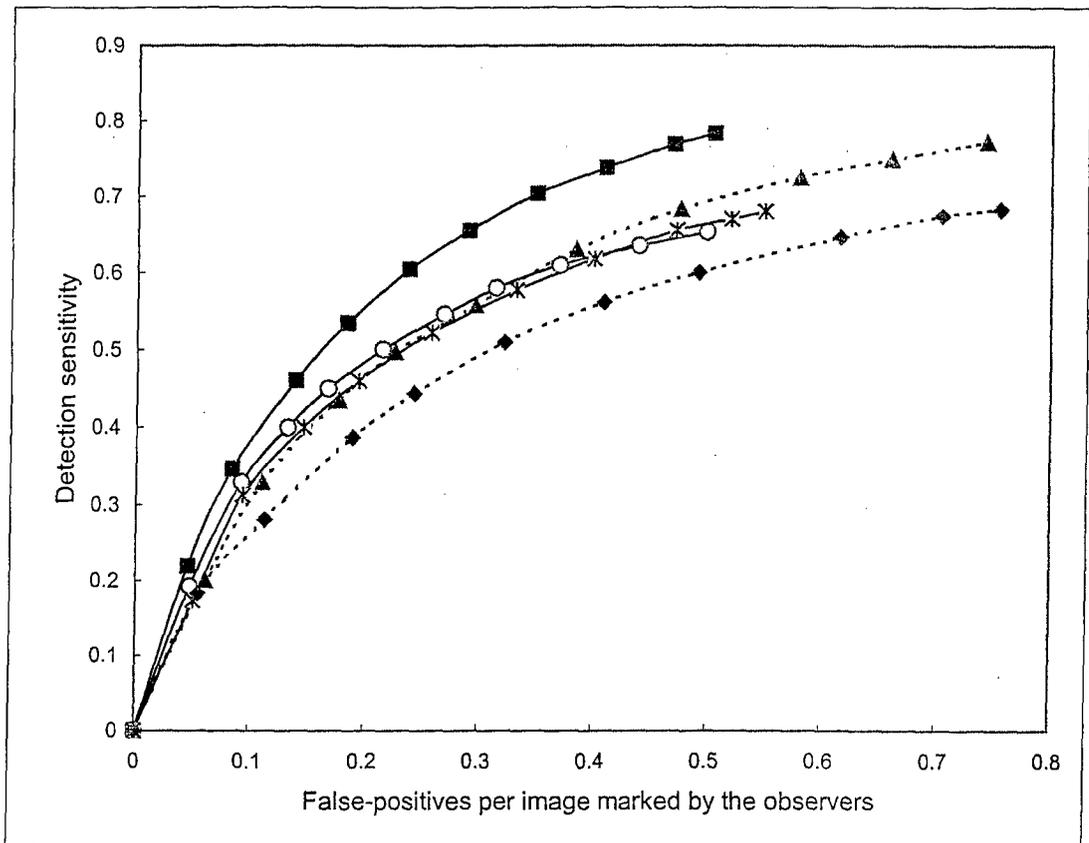


Figure 1: Curves of average detection performance of mammographic abnormalities (including both masses and microcalcification clusters) for seven participating radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (\*), and mode 5 (◆).

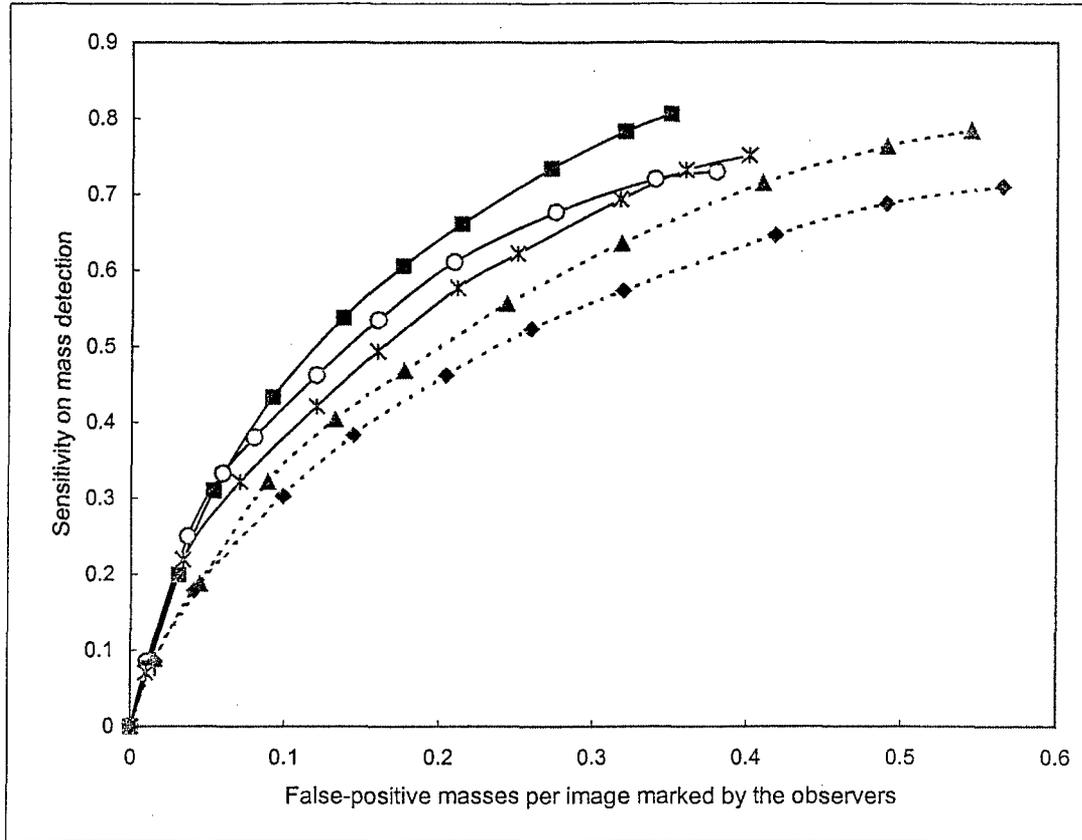


Figure 2: Curves of average performance of mass detection for seven radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (\*), and mode 5 (◆).

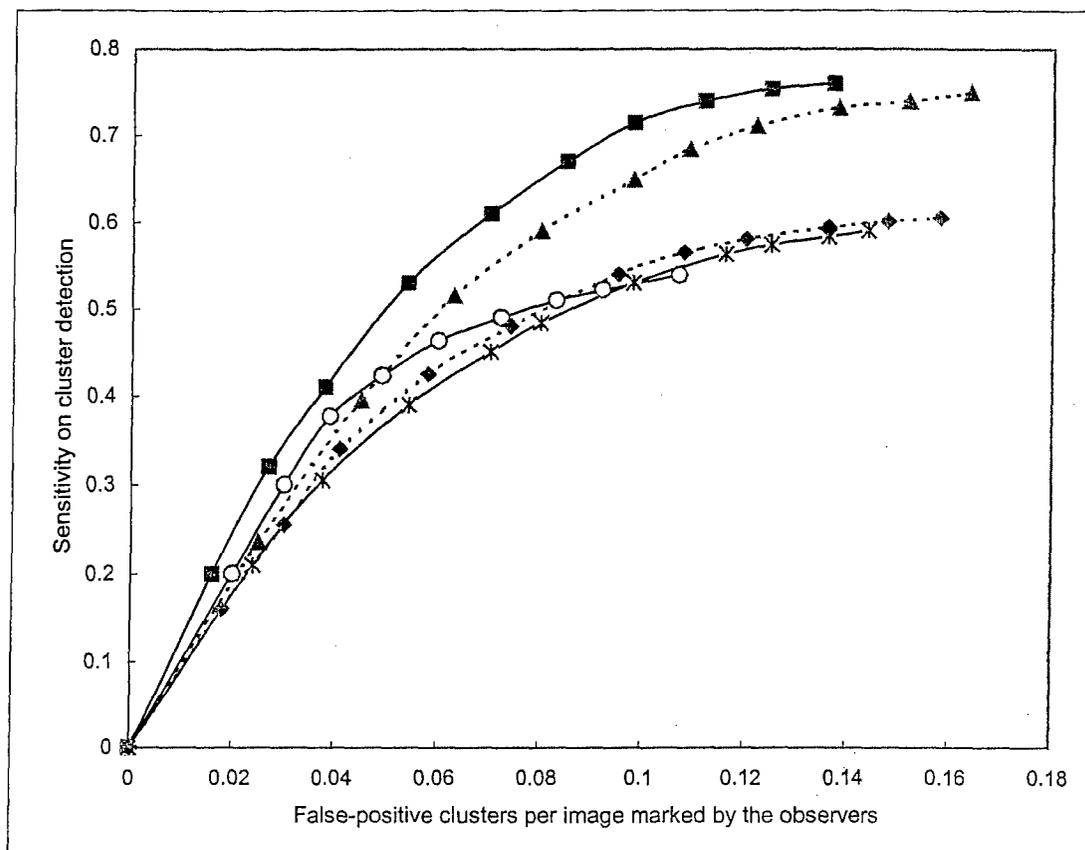


Figure 3: Curves of average performance of microcalcification cluster detection for seven radiologists using the five display modes. Display modes are represented as follows: mode 1 (o), mode 2 (■), mode 3 (▲), mode 4 (\*), and mode 5 (◆).

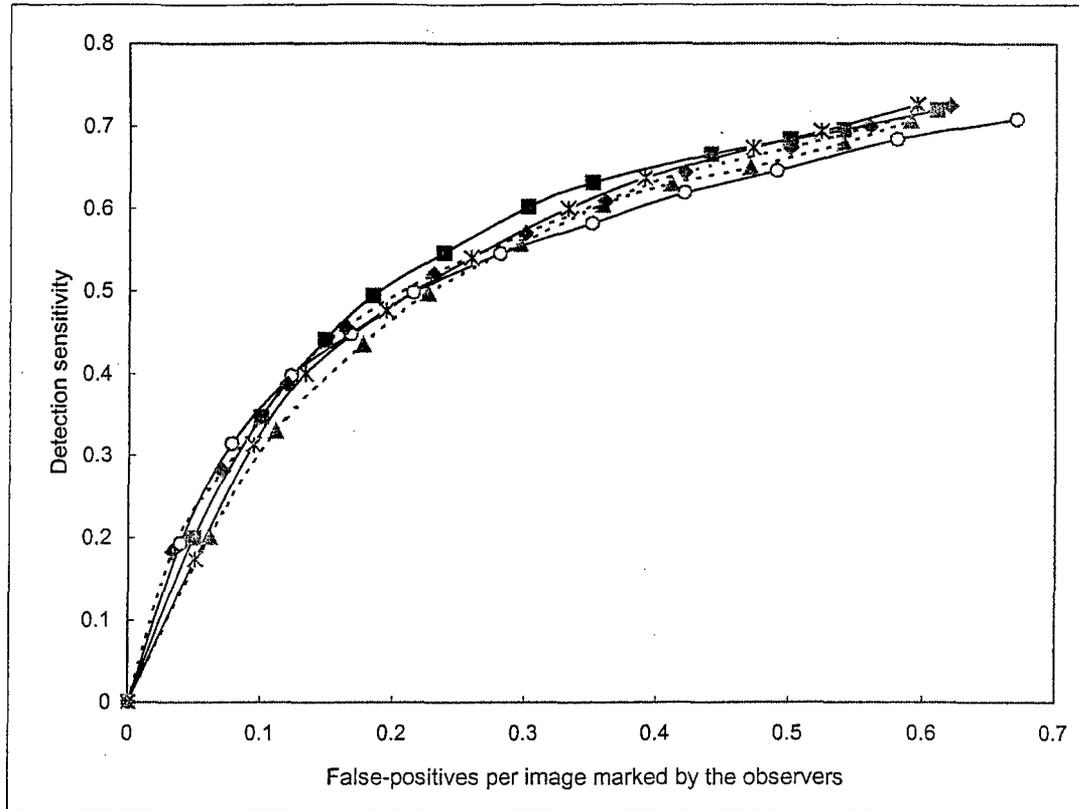


Figure 4: Curves of average detection performance of abnormalities for seven radiologists as a function of the order of appearance or round (e.g., first time, second time, etc), regardless of reading mode. Order of appearance is represented as follows: first time (o), second time (■), third time (▲), fourth time (\*), and fifth time (◆).



**DEPARTMENT OF THE ARMY**

US ARMY MEDICAL RESEARCH AND MATERIEL COMMAND  
504 SCOTT STREET  
FORT DETRICK, MARYLAND 21702-5012

REPLY TO  
ATTENTION OF:

MCMR-RMI-S (70-1y)

26 Aug 02

MEMORANDUM FOR Administrator, Defense Technical Information Center (DTIC-OCA), 8725 John J. Kingman Road, Fort Belvoir, VA 22060-6218

SUBJECT: Request Change in Distribution Statement

1. The U.S. Army Medical Research and Materiel Command has reexamined the need for the limitation assigned to technical reports written for this Command. Request the limited distribution statement for the enclosed accession numbers be changed to "Approved for public release; distribution unlimited." These reports should be released to the National Technical Information Service.

2. Point of contact for this request is Ms. Kristin Morrow at DSN 343-7327 or by e-mail at Kristin.Morrow@det.amedd.army.mil.

FOR THE COMMANDER:

Encl

  
PHYLLIS M. RINEHART  
Deputy Chief of Staff for  
Information Management

ADB274369  
ADB256383  
ADB264003  
ADB274462  
ADB266221  
ADB274470  
ADB266221  
ADB274464  
ADB259044  
ADB258808  
ADB266026  
ADB274658  
ADB258831  
ADB266077  
ADB274348  
ADB274273  
ADB258193  
ADB274516  
ADB259018  
ADB231912  
ADB244626  
ADB256677  
ADB229447  
ADB240218  
ADB258619  
ADB259398  
ADB275140  
ADB240473  
ADB254579  
ADB277040  
ADB249647  
ADB275184  
ADB259035  
ADB244774  
ADB258195  
ADB244675  
ADB257208  
ADB267108  
ADB244889  
ADB257384  
ADB270660  
ADB274493  
ADB261527  
ADB274286  
ADB274269  
ADB274592  
ADB274604

ADB274596  
ADB258952  
ADB265976  
ADB274350  
ADB274346  
ADB257408  
ADB274474  
ADB260285  
ADB274568  
ADB266076  
ADB274441  
ADB253499  
ADB274406  
ADB262090  
ADB261103  
ADB274372