AD_____

Award Number: DAMD17-00-1-0410

TITLE: Remote Patient Management in a Mammographic Screening Environment in Underserved Areas

PRINCIPAL INVESTIGATOR: David Gur, Sc.D.

CONTRACTING ORGANIZATION: University of Pittsburgh
Pittsburgh, PA 15260

REPORT DATE: September 2002

TYPE OF REPORT: Annual

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2002	Annual (1 Sep 01 - 31 Aug 02)

4. TITLE AND SUBTITLE
Remote Patient Management in a Mammographic Screening Environment in Underserved Areas

5. FUNDING NUMBERS
DAMD17-00-1-0410

6. AUTHOR(S)
David Gur, Sc.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pittsburgh
Pittsburgh, PA 15260

E-Mail: gurd@msx.upmc.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

Early detection of breast cancer is of significant interest to our society. Mammographic screening is gradually moving toward a "distributed acquisition – centralized review" approach. Unfortunately, a relatively high recall rate using this approach increases patient anxiety as well as the cost and complexity of the diagnostic process. The purpose of this project is to evaluate in a multi-phase project the possible impact of a unique telemammography system that utilizes common carriers with wavelet-based data compression for image transmission on the recall rate in remote locations where physicians are not available during mammographic procedures. The initial phase of the project encompasses the design, assembly, and technical testing of a multi-site telemammography system that enables the digitization, transmission, and display of wavelet compressed images as well as associated text documents of a case in less than 15 minutes. The impact of such a system with and without the incorporation of CAD results will be evaluated during a step-by-step assessment in a multi-site study.

14. SUBJECT TERMS
breast cancer, telemammography, detection, CAD

15. NUMBER OF PAGES
56

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Introduction

Periodic mass screening of asymptomatic women is rapidly gaining approval and acceptance, and the population segment recommended for screening is increasing due to both longer life expectancy as well as earlier recommended age for initial examination [1-3]. The large variability in a number of important aspects related to mammography, as practiced in the U.S., resulted in the enactment of the Mammography Quality Standards Act, which mandates accreditation of each program (facility, technical and professional) [4,5]. Shortages of expert mammographers in many locations, combined with the desire to make it convenient for the patient to undergo the procedure, suggest that there may be a need for high-quality telemammography systems that enable a distributed acquisition-centralized expert review type solution to the problem, particularly in underserved areas [6,7]. The relatively high recall rates (5-15%) of screened women to supplement information that was not ascertained during the initial visit (e.g. magnification views) also make it desirable to enable physician "monitoring" and "management" of remote locations so that patient-management decisions can be made while the patient remains in the clinic [8-11]. Current practices result in increased patient anxiety and added practice complexity and cost. Early attempts to develop and implement a practical telemammography solution to this problem failed due to several significant technical problems associated with acquisition, transmission, management, and display of the images [12-14]. Many of these technical issues have been resolved in recent years, but some remain [14-18]. Although an adequate communication infrastructure for high-quality telemammography is available within some urban regions, the fact remains that where it may be needed most (i.e. remote, non-urban locations), enabling (two-way) communication systems are limited mainly to the Plain Old Telephone System (POTS). Other communication technologies, such as satellites, are being evaluated for this purpose, but it is not likely that these will displace POTS in most underserved areas for quite some time [19-21]. Hence, the problem of cost effective, timely remote patient monitoring and management in many underserved areas is not a simple one.

As a part of this project, we are assembling and evaluating a unique telemammography system that enables improved communication between remote sites where physicians are not always available during the mammographic acquisition process and a central location where experts can review the acquired images shortly after acquisition and assess whether or not additional procedures (e.g., spot compression views) are needed [22,23]. The system we are assembling is based on prior preliminary experience acquired in our group during ten years of research in this general area. It includes the use of a common carrier for communication (Plain Old Telephone System, POTS), wavelet-based image compression for data reduction, and the optional incorporation of CAD results to the transmitted information. The main goal is to assess in a step-by-step approach whether the use of such a system could significantly reduce recall rates in the remote sites. Other secondary objectives regarding ways to improve communication and creating an environment for "more active" participation of the technologist in the diagnostic process are also being explored.

Body:

Since the initiation of the project on September 1, 2000, we have been progressing methodologically on the tasks listed in the Statement of Work (page 5 of the proposal), as

originally submitted. It should be noted that the project is, for the most part, back on track, schedule wise, despite the fact that the Imaging Research group was relocated during November and December 2000 from Scaife Hall of the University of Pittsburgh to Magee Womens Hospital of the University of Pittsburgh Medical Center Health System. While this move resulted in a minor interruption in adhering to the original schedule, in the long run, the project is benefiting from such a move, since the group is now located where much of the project is being carried out and evaluated. During year two of the project, work was performed in three different areas listed under Task 1 (Redesign and Assemble System), Task 2 (Implement Clinical System), and Task 3 (Clinical System's Evaluation) in the original proposal. We have also begun planning for Task 4.

Under Task 1, we performed the following:

With the exception of one task (1.c), which is partially completed (see comments below), we have completed all other tasks under this category. We assembled and tested a multi-site telemammography system that meets (and in some cases exceeded) our proposed specifications. The status of the tasks described under this category is as follows:

a) Select and Purchase Equipment: Completed.

b) Convert Software to Windows Based: Completed.

c) Develop Interface to FFDM Acquisition System: We enabled the system to accept DICOM images that will be required for completion of this task. However, due to strategic and market changes in this area (see special section below), we elected to postpone completion of this task.

As indicated in last year's report, we have acquired the GE FFDM system (not a part of the project) and obtained the information needed for transferring the DICOM images. Using our DICOM tool kit, we have been actually transferring FFDM images to a server. Hence, the capability to complete the interface has been verified, and much of the work needed for this task was carried out. However, the FFDM field has been progressing rapidly from an acquisition technology point of view (in that several companies are now offering high-quality systems), and the specific systems that may be ultimately implemented in the future in our remote sites have not been determined. As important, the cost associated with such implementation is quite high, and we are finding that in most remote sites (ours as well as others), there is a reluctance to move rapidly into digital acquisition (FFDM). Hence, as indicated in our previous report, we are ready to complete an interface to an FFDM system when it is deemed timely and appropriate. However, at this time, we continue to focus our efforts on film digitization. It is not clear that the use of FFDM devices in remote "underserved" sites for screening purposes is likely to be common or appropriate in the near future.

d) Develop a New User Interface for the Acquisition Sites: Completed and tested.

A remote site user interface was completed and tested, both subjectively and objectively (by sending over 100 cases through the system). After minor modifications that were based on users' comments, our data entry and case-sending routines were refined and finalized.

e) **Complete Data Compression Software Module: Completed and tested.**
A compression software scheme was finalized and tested. The scheme allows for a site-specific selectable level of compression to be used.

f) **Develop and Refine Measures of Image Fidelity that can be used to Automatically Monitor and Adjust (if needed) Compression Levels on an Image-by-Image Basis:** Based on two independent tests (see evaluation section below), at two compression levels, 50:1 and 75:1, we enabled a "dial-up" compression capability in the system. However, we are finding out that the high level of acceptance of either compression level practically eliminates the need for this option. Therefore, we are currently using the system with a fixed level of compression (75:1). We believe that we have achieved high-quality images at such high compression levels that second-order image-specific adjustments are not needed for all practical purposes.

g) **Integrate all Software Modules**: All software modules were successfully integrated.

h) **Develop Display Protocols for the Workstation**: User-friendly display protocols have been developed and tested extensively (see system evaluation section).

i) **Assemble System**: The system was assembled as proposed.

j) **Test System in Laboratory:** The system has been tested in the laboratory.

k) **Trouble Shoot, Refine, and Finalize System**: Through refinements, we increased the operational ease-of-use and reliability of the system and finalized the base configuration for implementation.

l) **Prepare Clinical Sites for Implementation**: All three remote sites were prepared for system implementation as required.

Under Task 2, we performed the following:

a) All needed equipment was moved to the appropriate locations at the three remote sites. At each location, the equipment (send station and digitizer) is located at an easily accessible place. At the central site, we placed the "receive" workstation in a "screening" reading room at a central location within the Breast Center. This required some construction that was completed at no cost to the project.

b) The complete system was reassembled on location.

c) Technical and operational performance levels were retested on site.

d) Different evaluation protocols for initial system evaluations were developed and implemented.

1) 100 cases were randomly selected at each site and transmitted to a central site to assess ease-of-use, reliability, reproducibility, and cycle times. The results clearly indicate

that cases from all sites at 15, 20, and 90 miles away can be transmitted with a full duty cycle time (from data entry at remote site to display) that easily meets our proposed specifications. A four-image case can be completed in less than seven minutes using 75:1 compression, which is less than half the time we originally specified.

2) We performed a multi-reader subjective assessment of image quality, and all participating radiologists rated the quality as acceptable or better for the task at hand.

3) We evaluated differences in image quality on film and soft display at zero (no), 50:1, and 75:1 compression ratios and found that only under extreme magnification, the 75:1 level can be identified (recognized), but image quality is not significantly degraded for all practical purposes.

Under Task 3, we performed the following:

a) **Collect Baseline Information Off Mode:** We continue to analyze the data available in our databases concerning patient distributions and process-related information. This includes the recall rate by physician, site, type, and reason for recall. We have also obtained patient satisfaction survey results as ascertained from internal and external surveys, which had been performed by our institution for other purposes outside this project. Last, we obtain records concerning the cycle time from the initial examination to a definitive diagnosis for cases that were not being recalled, as well as cases that were. This analysis is performed for the different sites in which we operate, including but not limited to the two Pittsburgh-region sites that are included in this project. This effort continues throughout the project as data are collected and analyzed regarding the above-mentioned variables. The effort described here is preliminary and will constitute the initial baseline (reference) information for comparison purposes.

b, c, and d) **Technical and Clinical System Evaluations:** We have begun to evaluate step-by-step the possible utility of the telemammography system. All of these studies have been performed in a historical prospective mode. The first study included the transmission of several hundred cases for measurements of system performance. These have been successfully completed, and our average cycle time is less than seven minutes for a four-image case. The second study included the review of 100 cases by three observers to subjectively assess image quality. The study indicated that radiologists feel "comfortable" to "extremely comfortable" to perform the tasks at hand. The third study included an assessment of the automatic setting of display parameters (Look-up-tables, "LUTs"). Two experienced observers rated 50 cases for this purpose, and the results clearly indicated that in most cases our default settings were, at a minimum, "acceptable." The observers' ratings were 2.64 ± 0.57 and 3.51 ± 0.53 on a 1-4 scale.

After minor refinements, we have been assessing the fraction of cases that are being manually adjusted, and we find that approximately 90% of cases are being viewed using the default settings, and only approximately 10% of the cases are manually adjusted. We believe that this is an excellent level of success in this difficult display environment.

Our first retrospective clinically simulated study included five radiologists rating 310 transmitted cases each. We evaluated their recommendations for the need of additional procedures. A total of 310 cases (4 images each) were reviewed without any additional information, such as history, prior reports, or prior images. The results are provided in the following table.

Reviewing and Rating Screening Mammography Exams, Telemammography Workstation Versus Clinical Interpretation

	Cases Actually Recalled (n = 310)	Agreement with cases actually recalled (n = 42)	Agreement with cases not recalled (n = 268)
Clinical interpretation	13.5% (42)	N/A	N/A
Radiologist 1	27.1% (84)	64.3% (27)	78.7% (211)
Radiologist 2	29.7% (92)	69.1% (29)	76.5% (205)
Radiologist 3	36.7% (114)	69.1% (29)	68.3% (183)
Radiologist 4	45.8% (142)	78.6% (33)	59.3% (159)
Radiologist 5	54.8% (170)	78.6% (33)	48.9% (131)
Radiologists Average	38.8% (120)	71.9% (30)	66.6% (178)

The table clearly indicates a significant "over read" by the radiologists when using the workstation. This over-reading level stemmed from three known reasons:

1. The awareness that this was a retrospective study that does not affect patient care during the workstation interpretation.
2. The assumption that in a study of this sort we would use an enriched set (typically 30-50% true positive cases) with subtle abnormalities; hence, they did not want to miss true-positive cases, resulting in a significant over-read.
3. The lack of additional information, such as prior reports or images, for comparison.

We are currently addressing the latter two issues, and we plan to perform a second study to verify that the over-read can be corrected.

e) Collecting Objective Performance Measures of Traditional Systems: We continue to record all objective performance measures using the conventional system (non-transmitted cases) as indicated before.

f) To date we used films only for evaluation of high levels of data compression, since all of our radiologists prefer to use the workstation for clinical review purposes. The use of films for selected difficult cases (in particular those with possible subtle microcalcification clusters) will be evaluated in the future. The FFDM interface was previously addressed.

As a result of our initial experiences, we are in the process of adding the following capabilities to the system and evaluating their impact on performance.

1) **Real-time "chat"** – To facilitate effective communication between the technologists in the remote sites and experienced radiologists, we have implemented a "chat" box type function. The chat box provides a real-time interactive capability. Chat boxes on both sides contain: patient demographics; message area; pull-down menus; and a free typing text area. Typical communication includes the technologist sending a chat dialog with each case indicating: breast, left or right; view, cradiocaudal and/or mediolateral oblique; finding, mass or calcifications; comparison with prior exam, baseline, new, or increased; and possible additional procedure, additional views and/or ultrasound. The radiologists reply after reviewing the case to do recommended procedure as suggested; no additional procedures are necessary; and do not do suggested procedure, but do X, Y, and Z. We are currently performing an "on-line" experiment to test the reliability and ease of use of this communication tool, which will be followed by a simulated clinical study.

2) **Case Folder Enables More than Four Images** – We are in the process of enabling the "case folder" to include scanned reports (text) as well as more than four images (e.g., prior examination). The former capabilities (scanned report) have been developed and are currently under testing. The latter will be implemented thereafter.

It should be noted that many of the undertakings (tasks) in this section are designed to be completed in a comprehensive, multi-step approach and will continue throughout the whole project.

Under Task 4, we performed the following:

a) CAD Implementation: Although this task is not scheduled for the second year, we completed the design of a modular software package that will enable the different CAD routines to be incorporated into the telemammography system at the remote (sending) sites and transmit the results to the central site. This task is planned for implementation in the middle of year three of the project. Testing will commence shortly thereafter. Since our CAD efforts continue to result in performance improvements, we intend to finalize the actual scheme to be integrated as late as possible for optimal performance. The system will be operational with and without CAD. Some CAD modules have already been developed and are currently being tested.

Key (Research) Accomplishments:

During the first two years of the project, we have been progressing according to the original plan and are addressing many of the technical tasks and operational issues associated

with the design and implementation of the multi-site telemammography system. The key accomplishments for the second year were:

- We implemented, tested, and installed a multi-site telemammography system that meets (and in some areas exceeds) the technical specifications we anticipated.
- We successfully transmitted over 1000 cases from three remote sites to the central site.
- We initiated a step-by-step comprehensive, technical, and clinical assessment protocol in a clinically simulated environment.
- We are refining the system based on preliminary results to address ease of use and utility issues in order to maximize the likelihood of success.
- We have been able to coherently engage a large team of administrative, technical, clinical (i.e., technologist), and physician personnel in a large and complicated project.

Reportable Outcomes:

The nature of this project is such that most of the work performed during the first two and one-half years of the project does not result in a significant reportable outcome. However, as we develop the system, several tasks are being performed where partial support (albeit quite limited) is provided by this project. For example, we are developing a software package to incorporate CAD results into the telemammography system during the third year of the project. The development of our CAD schemes continue, and the performance seems to be improving as we progress in optimizing step-by-step the various schemes we have developed. Therefore, several of our scientific reports acknowledge this project.

- Zheng B, Ganott MA, Britton CA, Hakim CM, Hardesty LA, Chang TS, Rockette HE, Gur D. Soft-copy mammographic readings with different computer-assisted diagnosis cuing environments: Preliminary findings. Radiology 2001; 221:663-640
- Zheng B, Chang Y-H, Good WF, Gur D. Performance gain in computer-assisted detection schemes by averaging scores generated from artificial neural networks with adaptive filtering. Med Phys 2001; 28: 2302-2308
- Drescher JM, Maitz GS, Leader JK, Sumkin JH, Poller WR, Klaman H, Zheng B, Gur D. Design considerations for a multi-site, POTS-based telemammography system. Proc SPIE 2002; 4685:416-421
- Zheng B, Shah R, Wallace L, Hakim C, Ganott MA, Gur D. Computer-aided detection in mammography: An assessment of performance on current and prior images. Acad Radiol 2002; in press
- Leader JK, Sumkin JH, Drescher JM, Maitz GS, Zheng B, Wallace L, Hakim C, Hertzberg TM, Hardesty L, Shah R, Clearfield R, Sneddon C, Lindeman S, Craig D, Pugliese F, Duffner D, Lockhart J, Traylor C, Gur D. A multi-site telemammography system: technical challenges, operational issues, and preliminary clinical evaluation. To be presented at the Department of Defense "Era of Hope" meeting, September 25, 2002.

We anticipate that some of the design parameters and image testing will be reported at upcoming national meetings (e.g., SPIE).

Conclusions:

There are several technical, clinical, and assessment tasks listed in the Statement of Work of this project. During the first two years, we have addressed many technical tasks associated with the design and implementation of a multi-site telemammography system. We overcame many of the technical problems and assembled a multi-site system that exceeds several of the performance goals we originally proposed. The system is undergoing a comprehensive step-by-step evaluation (and refinement as deemed appropriate), and the goal is to establish and test an environment with improved communications capabilities between remote (and often underserved) facilities and a central site.

So What?

The main goal of this project is to evaluate how the use of an "almost real-time" telemammography system (with and without the use of CAD results) may impact the diagnostic process in terms of complete cycle time and patients' recall rate. At this stage, when we focus on system implementation and initial evaluations, it is premature to consider any impact statements that are relevant to the clinical environment. The nature of this project necessitates that the clinical evaluation requires a multi-step approach, hence, actual clinical results can only be provided at a later date. Success of this project will enable a demonstration and careful assessment of different ways to increase communication between remote (and potentially underserved sites) and a central site. Our hope is that by using this approach, one may be able to provide better, more cost-effective service at these sites, and reduce recall rates in these facilities by a significant amount.

References:

1. S Pelikan, M Moskowitz, "Effects of lead time length bias, and false-negative assurance on screening for breast cancer," Cancer 71, 1998-2005 (1993).
2. L Tabar, G Fagerberg, HH Chen, SW Duffy, CR Smart, A Gad, RA Smith, "Efficacy of breast cancer screening by age: New results from the Swedish Two-Country Trial," Cancer **75**, 2507-2517 (1995).
3. F Houn, ML Brown, "Current practice of screening mammography in the United States: Data from the national survey of mammography facilities," Radiology **190,** 209-215 (1994).
4. CA Beam, PM Layde, DC Sullivan, "Variability in the interpretation of screening mammograms by US radiologists," Arch Intern Med **156**, 209-213 (1996).

5. Food and Drug Administration, "Mammography facilities: requirements for accrediting bodies and quality standards and certification requirements-interim rules," Federal Register **58**, 67558-72. (CFR21, Part 900) (1993).
6. JG Elmore, CK Wells, CH Lee, DH Howard, AR Feinstein, "Variability in radiologists' interpretations of mammograms," N Engl J Med **331**, 1493-1499 (1994).
7. RML Warren, SW Duffy, "Comparison of single reading with double reading of mammograms and change in effectiveness with experience," Br J Radiol **68**, 958-962 (1995).
8. CJ Wright, CB Mueller, "Screening mammography and public health policy: The need for perspective," Lancet **346**, 29-32 (1995).
9. LW Bassett, RE Hendrick, TL Bassford, PF Butler, D Carter, M DeBor, CJ D'Orsi, CJ Garlinghouse, RF Jones, AS Langer, JL Lichtenfeld, JR Osuch, LN Reynolds, ES de Paredes, RE Williams, "Responsibilities of the mammography facility," In: Quality determinants of mammography, clinical practice guideline. Number 13. Washington, DC: US Department of Health and Human Services, AHCPR publication no. 95-0632 (1994).
10. JG Elmore, MB Barton, VM Moceri, S Polk, PJ Arena, SW Fletcher, "Ten-year risk of false-positive screening mammograms and clinical breast examinations," N Engl J Med **338**, 1089-1096 (1998).
11. DS May, NC Lee, MR Nadel, RM Henson, DS Miller, "The National Breast and Cervical Cancer Early Detection Program: Report of the First 4 Years of Mammography Provided to Medically Underserved Women," AJR **170**, 97-104 (1998).
12. SA Feig, MJ Yaffe, "Digital mammography, computer-aided diagnosis, and telemammography," Radiol Clin North Am **33**, 1205-1228 (1995).
13. LL Fajardo, MT Yoshino, GW Seeley, R Hunt, TB Hunter, R Friedman, D Cardenas, R Boyle, "Detection of breast abnormalities on teleradiology transmitted mammograms," Invest Radiol **25**, 1111-1115 (1990).
14. MA Goldberg, "Telemammography: Implementation issues," Telemedicine Journal **1,** 215-226 (1995).
15. HK Huang, SL Lou, E Sickles, D Hoogstrate, M Jahangiri, F Cao, J Wang, "Technical issues in full-field direct digital telemammography," [Chapter] In: Computer Assisted Radiology and Surgery. Lemke HU, Inamura K, Editors. Elsevier Science B.V., 662-667 (1997).
16. HK Huang, "Digital Mammography: A Missing Link in a Totally Digital Radiology Department," Presented at the EuroPACS 97 Meeting; PISA, Italy. September 25-27, (1997).
17. JM Murphy, NJ O'Hare, D Wheat, PA McCarthy, A Dowling, R Hayes, H Bowmer, GF Wilson, MP Molloy, "Digitized mammograms: a preliminary clinical evaluation and the potential for telemammography," Journal of Telemedicine and Telecare **5,** 193-197 (1999).
18. SL Lou, HD Lin, KP Lin, D Hoogstrate, "Automatic breast region extraction from digital mammograms for PACS and telemammography applications," Computerized Medical Imaging and Graphics **24**, 205-220 (2000).
19. S Dwyer, Private communications. See also "Telemedicine Targets Mammographic Services" in Biophotonics International Nov/Dec 1997. Page 10.
20. SL Lou, EA Sickles, HK Huang, D Hoogstrate, F Cao, J Wang, M Jahangiri, "Full-field direct digital telemammography: Technical components, study protocols, and preliminary results," IEEE Trans Info Technology in Biomedicine **1,** 270-278 (1997).

21. SL Lou, HK Huang, E Sickles, D Hoogstrate, F Cao, J Wang, "Full-field direct digital telemammography: system implementation," Proc SPIE **3339**, 156-164 (1998).
22. GS Maitz, TS Chang, JH Sumkin, PW Wintz, CM Johns, M Ganott, BL Holbert, CM Hakim, KM Harris, D Gur, JM Herron, "Preliminary clinical evaluation of a high-resolution telemammography System," Invest Radiol **32**, 236-240 (1997).
23. JM Holbert, M Staiger, TS Chang, JD Towers, CA Britton, "Selection of processing algorithms for digital image compression: A rank-order study," Acad Radiol **2,** 273-276 (1995).

PROGRESS IN BIOMEDICAL OPTICS AND IMAGING

Reprinted from

Medical Imaging 2002

PACS and Integrated Medical Information Systems: Design and Evaluation

26–28 February 2002
San Diego, USA

Proceedings of SPIE
Volume 4685

Design Considerations for a Multi-Site, POTS-Based Telemammography System

John M. Drescher*, Glenn S. Maitz, J. Ken Leader, Jules H. Sumkin, William R. Poller, Herta Klaman, Bin Zheng, David Gur

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

ABSTRACT

As the number of mammographic examinations increases, it becomes clear that in many underserved locations, there is a lack of expertise that is required for consistent, highly accurate, and timely diagnosis. Hence, mammograms are frequently sent to other medical facilities, and a significant fraction of women (typically 3-10%) are recalled for additional examinations. It is the purpose of this project to develop, test, and clinically evaluate a telemammography system that will operate between several remote locations and a large breast cancer center. In this manuscript we describe the design considerations, implementation, and initial testing that were undertaken, to date. The system digitizes a mammogram at 50 μm pixel size, compresses the resulting image file (~75:1), and transmits it over a telephone line to the central site where the data received are decompressed and displayed on a high-resolution workstation in approximately 4 minutes per image. Initial testing of the system indicates that a relatively inexpensive system for "almost real-time" telemammography can be employed in any geographic area that possesses standard telephone lines, and this approach to enhance communication may make it possible to offer better mammographic services at remote locations.

Key Words: Imaging, Teleradiology, Mammography, Data compression, Image display

1. INTRODUCTION

Periodic mass screening of asymptomatic women is rapidly gaining approval and acceptance, and the population segment recommended for screening is increasing due to both longer life expectancy as well as earlier recommended age for initial examination [1-3]. The large variability in a number of important aspects related to mammography, as practiced in the U.S., resulted in the enactment of the Mammography Quality Standards Act, which mandates accreditation of each program (facility, technical and professional) [4,5]. Shortages of expert mammographers in many locations, combined with the desire to make it convenient for the patient to undergo the procedure, suggest that there may be a need for high-quality telemammography systems that enable a distributed acquisition-centralized expert review type solution to the problem [6,7]. The relatively high recall rates (5-15%) of screened women to supplement information that was not ascertained during the initial visit (e.g. magnification views) also make it desirable to enable physician "monitoring" and "management" of remote locations so that clinical and diagnostic decisions can be made while the patient remains in the clinic [8-11]. Early attempts to develop and implement a practical telemammography solution to this problem failed due to several significant technical problems associated with acquisition, transmission, management, and display of the images [12-14]. Many of these technical issues have been resolved in recent years, but some remain [14-18]. Although an adequate communication infrastructure for high-quality telemammography is available within some urban regions, the fact remains that where it may be needed most (i.e. remote, non-urban locations), enabling (two-way) communication systems are limited mainly to the Plain Old Telephone System (POTS). Other communication technologies, such as satellites, are being evaluated for this purpose, but it is not likely that these

*jdrescher@mail.magee.edu; phone 412-641-2563; fax 412-641-2582://www.radiology.upmc.edu/University of Pittsburgh, Suite 4200 Magee Womens Hospital, 300 Halket Street, Pittsburgh, PA, 15213, USA

Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation, Eliot L. Siegel, H. K. Huang, Editors, Proceedings of SPIE Vol. 4685 (2002) © 2002 SPIE · 1605-7422/02/$15.00

will displace POTS in most underserved areas for quite some time [19-21]. Hence, the problem of cost effective, timely remote patient monitoring and management in many underserved areas is not a simple one. Using a unique data-handling scheme, we have been able to demonstrate that high-quality, multi-site telemammography systems can be developed under these acquisition and communication constraints [22,23]. Using similar concepts, we have been developing a multi-site system that enables "almost real-time communications" between the "spokes and the hub." Design considerations as well as implementations and initial testing procedures are described in the manuscript.

2. METHOD

At the remote sites, we use a high resolution Lumiscan 85 film digitizer (Eastman Kodak, Rochester, NY) connected via SCSI to a Windows NT 2000 PC (900 MHz Athlon 512 MB) running multi-threaded software. The digitizer is equipped with a film feeder and is capable of digitizing up to six films in a batch at 50 μm pixel size over optical densities ranging from 0 to 4.0 OD. Four slots of the film feeder are labeled for specific mammographic views (i.e. LCC, RCC, RMLO and LMLO) for ease of use during the digitization process. The user at the remote site (typically a technologist) selects either an option to digitize a "standard" protocol for an image set or any of the six films he/she chooses to send, by clicking on an appropriate icon.

The user enters patient information into a computer data entry form during the digitization. At this time he/she also enters information for 'non-standard' cases by choosing from drop-down menus the anatomy and view for each of the films being digitized. Meanwhile the software on the PC establishes a connection with the central hub if a connection does not already exist. This is currently done via dial-up phone line or an Internet connection, but optionally ISDN or DSL can be used as well. For the dial-up connection, internal 56K hardware modems (U.S. Robotics, Rolling Meadows, IL) are used. The image data are processed in sections, segmented, and compressed using JPEG 2000 compatible irreversible wavelet compression and transmitted in packets to the central site. Optionally, a report or patient history can be transmitted along with the images by inserting them into an attached page scanner (OneTouch 8650, Visioneer, Inc., Fremont, CA).

The central site has a Windows 2000 Server workstation (Dual 1.2 GHz Athlon MP, 2 GB RAM) running specially developed software. Data received from remote sites is reconstructed from the packets, decompressed, and stored on a hard disk and/or in memory (if available). Several cases (depending on size) can be stored in memory for instant access. Cases stored on disk take a few seconds to restore to memory. The display consists of a pair of high-resolution (2048 x 2560) 8-bit grayscale portrait monitors at a nominal setting of 80 ftL (DS5100P, Clinton Electronics, Rockford, IL). The bottom of the displays holds a bar of icons and arrows for selecting cases, images, and other tools. The user can select from a patient list that displays the unreviewed cases on the top (similar to a "worklist"). When a case is selected, four images appear in quadrants on the right monitor. The left monitor displays the currently selected image (the first image by default) at half the available resolution. Although images are displayed at window and level settings determined by the statistics of the signal from individual image data sets, the user may select the window and level tool and alter it in real time using a mouse. A "magnify" tool is also available that magnifies any square region under the cursor in real-time to full resolution as it is moved over the image. Among other tools on the tool bar are arrows that allow movement to the next or preceding case.

We plan to add DICOM compatibility to the workstation at the central site. This will include the capability to send and print selected images to a mammographic film printer (DryView 8610, Eastman Kodak, Rochester, NY). This will also allow transferring workstation images to another DICOM device (workstation or storage) and also allow access to images from other DICOM compatible devices, such as full field digital mammography acquisition systems [24,25].

We also plan to add computer-aided detection (CAD) software at the remote site. This would allow image analysis to be performed on the original images during the time the compressed data sets are transmitted. The results can be sent immediately after the image data transfer, simply as coordinate data. Suspicious areas for masses and microcalcifications would then be marked on a removable overlay on the images at the hub. Figure 1 is a schematic diagram of the system as it is currently configured and being evaluated.

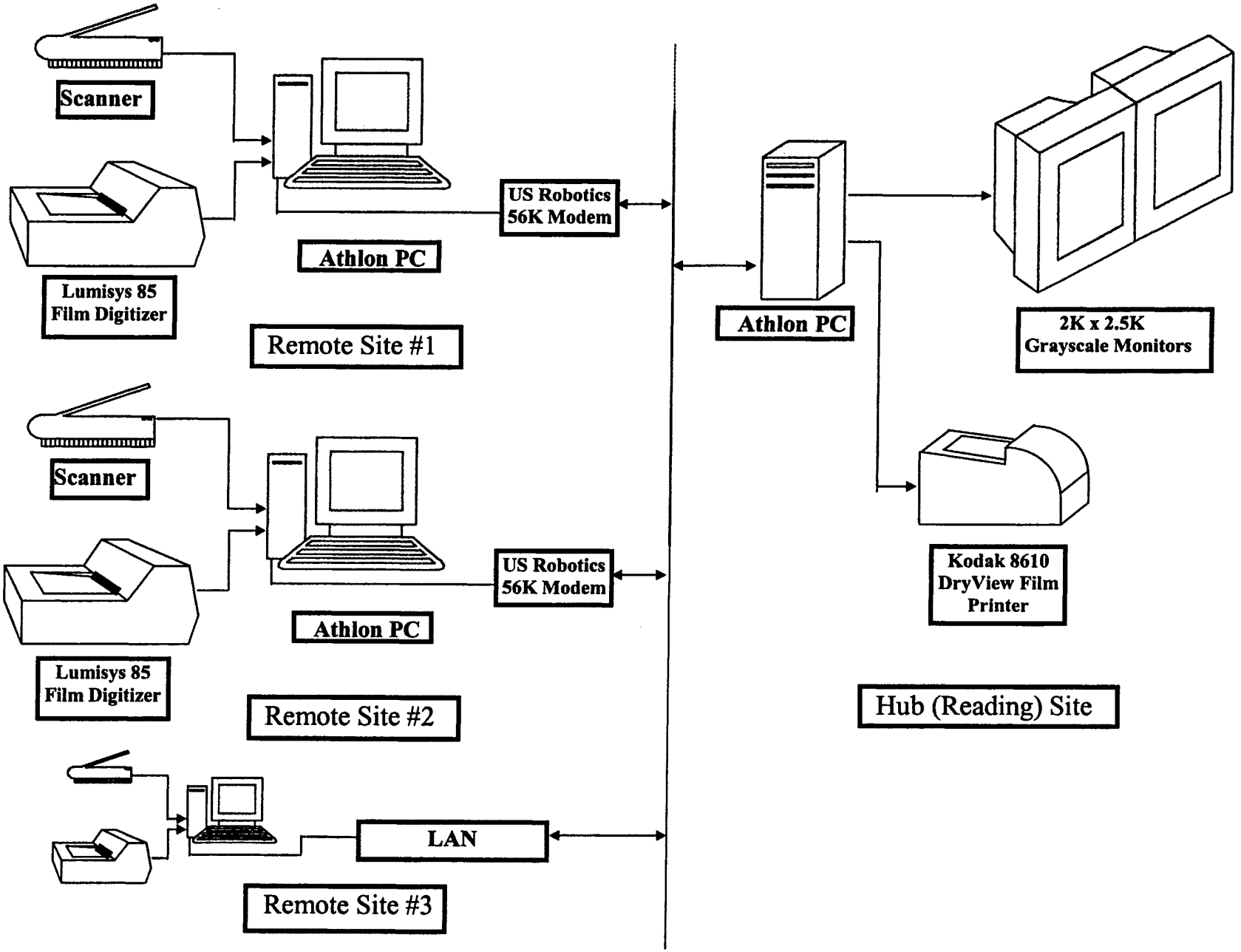

Figure 1: Schematic of the Telemammography System

3. SOFTWARE DESIGN

Both the hub site and remote site computer programs are designed using multithreading to permit each task to be completed in a timely manner; yet, allow the system to be responsive to user input. The main threads communicate with one other by sending thread messages to other threads. Each main thread handles the messages that are applicable to it and ignores any others. A main thread may spawn another thread to accomplish some subordinate task. These spawned threads do not receive messages, but they do send messages.

The main threads for the hub site program are:
Archive Manager that handles saving and loading of images and cases and the deletion of uncompressed images when free disk space becomes low.
The Case Manager handles the functions of creating images and cases, in addition to most of the database functions.
The Display Manager controls the display of images and forwards messages to the main application window.
The Distribution Manager handles the receipt and transmission of data and the processing (including decompression) of the data.

The main threads for a remote site are:
Digitizer Manager that handles all the tasks related to the film digitization.
The Case Manager handles the functions of creating images and cases, in addition to most of the database functions.
The Display Manager controls the display of images and forwards messages to the main application window.
The Distribution Manager handles the transmission and receipt of data and the processing (including compression) of the data.

The threads for the most part are synchronized using a Reader / Writer lock that is a combination of the built-in Microsoft Windows synchronization primitives. This lock allows either any number of readers or just one writer to have access to a shared object. This allows greater concurrency than that which could be achieved by using a Mutex, which allows only a single thread to access an object at a time forcing all other threads to wait.

4. USER FUNCTIONALITY

At the remote sites all data entry functions utilize pull-down menus supported by the use of a keyboard. A "start" command enables digitization of a case, and data entry can be performed within a predetermined time slot during the digitization process. At the central site, a high-resolution workstation is operated solely using a mouse, and several simple options are available by clicking on the appropriate button (e.g. flip, magnify, rotate, display on other monitor, etc.). The cases in memory and those on disk are so indicated on patient lists, and automatic lookup tables (image-statistic based) are used to display "reasonable" default settings.

5. RESULTS

The system has been designed, assembled and tested for technical reliability. Currently the three sites (See Figure 1) are located anywhere from 15-90 miles away from our hub in Pittsburgh. The remote sites are all outpatient clinics, which are staffed by a physician between one day a week to half a day every two weeks. Cases from multiple sites have been transmitted simultaneously and received successfully at the hub. Average transmission times for a four-image case vary significantly based on bandwidth availability and film size and currently ranges from 9 to 25 minutes. We are currently evaluating different approaches to reduce the cycle time to below 15 minutes per case as an upper limit. To date we have received over 200 cases from the remote sites, and we are analyzing user functionality at all locations.

Two mammographers performed an initial evaluation of a series of cases and the basic workstation's basic functionality. The quality of the images received was subjectively judged to be acceptable or better. A series of retrospective analyses on a large number of cases sent from all sites will follow.

6. DISCUSSION

Low cost telemammography is becoming feasible as communication technology and processing capabilities continue to improve in terms of cost, availability, and reliability. The system we designed is capable of variable compression rates, should it be desired, as well as the ability to print images at the receiving site. As important, the incorporation of a CAD scheme into the protocol may aid in decision making at both the sending (remote) sites as well as the receiving site. It should be noted that the system was not designed for electronic primary diagnosis, but rather to facilitate better communication between remote (and perhaps underserved) sites and a central hub where expertise is more readily available.

Our initial assessment indicates that technically our objectives can be met, and we hope that our planned clinical evaluations will improve our understanding as to whether or not such systems can be used to enhance communication, aid in timely decision making, help reduce recall rates, and ultimately enhance and improve the timeliness and quality of the service we can provide in locations where expert mammographers are not physically present at the time of the examination.

ACKNOWLDEGEMENTS

This work is supported in part by grant DAMD17-00-1-0410 from the Department of Defense. 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.

REFERENCES

1. S Pelikan, M Moskowitz, "Effects of lead time length bias, and false-negative assurance on screening for breast cancer," Cancer 71, 1998-2005 (1993).
2. L Tabar, G Fagerberg, HH Chen, SW Duffy, CR Smart, A Gad, RA Smith, "Efficacy of breast cancer screening by age: New results from the Swedish Two-Country Trial," Cancer **75**, 2507-2517 (1995).
3. F Houn, ML Brown, "Current practice of screening mammography in the United States: Data from the national survey of mammography facilities," Radiology **190**, 209-215 (1994).
4. CA Beam, PM Layde, DC Sullivan, "Variability in the interpretation of screening mammograms by US radiologists," Arch Intern Med **156**, 209-213 (1996).
5. Food and Drug Administration, "Mammography facilities: requirements for accrediting bodies and quality standards and certification requirements-interim rules," Federal Register **58**, 67558-72. (CFR21, Part 900) (1993).
6. JG Elmore, CK Wells, CH Lee, DH Howard, AR Feinstein, "Variability in radiologists' interpretations of mammograms," N Engl J Med **331**, 1493-1499 (1994).
7. RML Warren, SW Duffy, "Comparison of single reading with double reading of mammograms and change in effectiveness with experience," Br J Radiol **68**, 958-962 (1995).
8. CJ Wright, CB Mueller, "Screening mammography and public health policy: The need for perspective," Lancet **346**, 29-32 (1995).
9. LW Bassett, RE Hendrick, TL Bassford, PF Butler, D Carter, M DeBor, CJ D'Orsi, CJ Garlinghouse, RF Jones, AS Langer, JL Lichtenfeld, JR Osuch, LN Reynolds, ES de Paredes, RE Williams, "Responsibilities of the mammography facility," In: Quality determinants of mammography, clinical practice guideline. Number 13. Washington, DC: US Department of Health and Human Services, AHCPR publication no. 95-0632 (1994).
10. JG Elmore, MB Barton, VM Moceri, S Polk, PJ Arena, SW Fletcher, "Ten-year risk of false-positive screening mammograms and clinical breast examinations," N Engl J Med **338**, 1089-1096 (1998).
11. DS May, NC Lee, MR Nadel, RM Henson, DS Miller, "The National Breast and Cervical Cancer Early Detection Program: Report of the First 4 Years of Mammography Provided to Medically Underserved Women," AJR **170**, 97-104 (1998).
12. SA Feig, MJ Yaffe, "Digital mammography, computer-aided diagnosis, and telemammography," Radiol Clin North Am **33**, 1205-1228 (1995).

13. LL Fajardo, MT Yoshino, GW Seeley, R Hunt, TB Hunter, R Friedman, D Cardenas, R Boyle, "Detection of breast abnormalities on teleradiology transmitted mammograms," Invest Radiol **25**, 1111-1115 (1990).
14. MA Goldberg, "Telemammography: Implementation issues," Telemedicine Journal **1**, 215-226 (1995).
15. HK Huang, SL Lou, E Sickles, D Hoogstrate, M Jahangiri, F Cao, J Wang, "Technical issues in full-field direct digital telemammography," [Chapter] In: Computer Assisted Radiology and Surgery. Lemke HU, Inamura K, Editors. Elsevier Science B.V., 662-667 (1997).
16. HK Huang, "Digital Mammography: A Missing Link in a Totally Digital Radiology Department," Presented at the EuroPACS 97 Meeting; PISA, Italy. September 25-27, (1997).
17. JM Murphy, NJ O'Hare, D Wheat, PA McCarthy, A Dowling, R Hayes, H Bowmer, GF Wilson, MP Molloy, "Digitized mammograms: a preliminary clinical evaluation and the potential for telemammography," Journal of Telemedicine and Telecare **5**, 193-197 (1999).
18. SL Lou, HD Lin, KP Lin, D Hoogstrate, "Automatic breast region extraction from digital mammograms for PACS and telemammography applications," Computerized Medical Imaging and Graphics **24**, 205-220 (2000).
19. S Dwyer, Private communications. See also "Telemedicine Targets Mammographic Services" in Biophotonics International Nov/Dec 1997. Page 10.
20. SL Lou, EA Sickles, HK Huang, D Hoogstrate, F Cao, J Wang, M Jahangiri, "Full-field direct digital telemammography: Technical components, study protocols, and preliminary results," IEEE Trans Info Technology in Biomedicine **1**, 270-278 (1997).
21. SL Lou, HK Huang, E Sickles, D Hoogstrate, F Cao, J Wang, "Full-field direct digital telemammography: system implementation," Proc SPIE **3339**, 156-164 (1998).
22. GS Maitz, TS Chang, JH Sumkin, PW Wintz, CM Johns, M Ganott, BL Holbert, CM Hakim, KM Harris, D Gur, JM Herron, "Preliminary clinical evaluation of a high-resolution telemammography System," Invest Radiol **32**, 236-240 (1997).
23. JM Holbert, M Staiger, TS Chang, JD Towers, CA Britton, "Selection of processing algorithms for digital image compression: A rank-order study," Acad Radiol **2**, 273-276 (1995).
24. S Muller, "Full-field digital mammography designed as a complete system," European Journal of Radiology **31**, 25-34 (1999).
25. JM Lewin, RE Hendrick, CJ D'Orsi, PK Isaacs, LJ Moss, A Karellas, GA Sisney, CC Kuni, GR Cutter, "Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations," Radiology **218**, 873-880 (2001).

Performance gain in computer-assisted detection schemes by averaging scores generated from artificial neural networks with adaptive filtering

Bin Zheng,[a] Yuan-Hsiang Chang, Walter F. Good, and David Gur
Department of Radiology, University of Pittsburgh, Pittsburgh, Pennsylvania 15213

(Received 27 February 2001; accepted for publication 27 August 2001)

The authors investigated a new method to optimize artificial neural networks (ANNs) with adaptive filtering used in computer-assisted detection schemes in digitized mammograms and to assess performance changes when averaging classification scores from three sets of optimized schemes. Two independent training and testing image databases involving 978 and 830 digitized mammograms, respectively, were used in this study. In the training data set, initial filtering and subtraction resulted in the identification of 592 mass regions and 3790 suspicious, but actually negative regions. These regions (including both true-positive and negative regions) were segmented into three subsets three times based on the calculation of the values of three features as segmentation indices. The indices were "mass" size multiplied by their digital value contrast, conspicuity, and circularity. Nine ANN-based classifiers were separately optimized using a genetic algorithm for each subset of regions. Each region was assigned three classification scores after applying the three adaptive ANNs. The performance gain of the CAD scheme after averaging the three scores for each suspicious region was tested using an independent data set and a ROC methodology. The experimental results showed that the areas under ROC curves (A_z) for the testing database using three sets of optimized ANNs individually were 0.84 ± 0.01, 0.83 ± 0.01, and 0.84 ± 0.01, respectively. The between-index correlations of three A_z values were 0.013, -0.007, and 0.086. Similar to averaging diagnostic ratings from independent observers, by averaging three ANN-generated scores for each testing region, the performance of the CAD scheme was significantly improved ($p<0.001$) with A_z value of 0.95 ± 0.01.
[DOI: 10.1118/1.1412240]

Key words: computer-assisted diagnosis, mammography, mass detection, artificial neural network, genetic algorithm, adaptive filtering

I. INTRODUCTION

A number of computer-assisted detection (CAD) schemes have been developed in recent years to detect masses and microcalcification clusters depicted in digitized mammograms.[1–10] Many researchers believe that eventually these CAD schemes will help radiologists to significantly improve their diagnostic accuracy and efficiency in diagnosing breast cancers at an earlier stage.[11–13] Others question whether the high false-positive rates resulting from the CAD schemes could generate a large number of unnecessary recalls or possibly biopsies, which might offset the possible gains in detection sensitivity.[14,15] Because of this potential negative effect (i.e., high false-positive rate) on diagnostic performance, significant effort has been invested in an attempt to improve CAD performance.[16–19] In order to achieve high detection sensitivity, CAD schemes typically identify a large number of suspicious, but actually negative regions at the initial detection stage. Hence, an important task in CAD development is to improve accuracy of classifying a large number of identified regions. Previous studies in this area focused mainly on searching for an effective classifier including, but not limited to: a linear discriminant function,[5] an improved artificial neural network (ANN),[20] a wavelet transformation,[3] a set enumeration decision tree,[21] a Bayesian belief network,[22] and a knowledge-based expert system.[23] Other efforts concentrated on determining a small, but optimal set of features that include morphological features,[10] texture features,[16] and derivative-based features.[4]

Because of the complexity and large variability of the abnormalities in question and the surrounding tissue structures, it is quite difficult for a single universal scheme to accurately classify suspicious regions using a limited number of correlated features.[24,25] To address this problem, two approaches have been investigated to date. The first one is to segment the images or suspicious regions into different groups based on specific predetermined image characteristics (e.g., "image difficulty indices") and then optimize separate schemes with adaptive filtering for each group (class) of images. Previous studies using this approach suggested promising results for a rule-based CAD scheme[26] and for a wavelet-transform based CAD scheme.[27] The second approach that has been explored is to combine (or average) the detection results from different noncorrelated classifiers, such as the averaging of detection scores from a rule-based and ANN-based classifiers,[17] or those of an ANN and a set enumeration tree.[21] Similar to improving diagnostic accuracy by averaging ratings from replicated, but independent read-

ings or from different readers,[28,29] averaging CAD scores generated by different classifiers could also be an effective approach to improve performance.[17,21]

In our previously reported studies,[21,26] image databases were somewhat limited and the computation of the indices by which images were segmented into groups was quite complicated. In the present study, we combine the two approaches. In addition, we use three image features that are well defined, easily computable, and widely used in CAD schemes to segment the image ensemble into different groups. This study focuses on detecting masses in digitized mammograms. Since studies have shown that high-performing CAD cueing could significantly improve the performance of radiologists in detecting subtle cancers[13,30–32] and our study suggested that once detected, the task of classifying masses as benign or malignant was not affected by the CAD detection performance, we assume here that detection and classification are two distinct and largely independent tasks.[32] A detailed description of the development phase of the scheme and the initial test using a large independent data set are presented.

II. MATERIALS AND METHODS

A. Image databases

Two independent image databases were used in this study. The first database (used as the training database) contains a total of 978 digitized mammograms. Of these, 545 images were acquired on patients who underwent mammographic examinations at the University of Pittsburgh Medical Center (Pittsburgh, PA) and its affiliated hospitals and clinics prior to April 1997, and 433 images were provided to us by an imaging research group at Washington University Medical School (St. Louis, MO). A detailed description of this database has been reported elsewhere.[22] The second image database (used as the testing database) contains 830 images, of which 528 were provided to us by a research and development team at the Eastman Kodak Company (Rochester, NY)[10] and 302 images collected more recently (>10/98) on patients undergoing mammography examinations at the University of Pittsburgh Medical Center. Although the mammograms originated in different medical facilitates, these were all digitized in our laboratory using a laser-film digitizer (Lumisys, Sunnyvale, CA) with a pixel size of $100\,\mu m \times 100\,\mu m$ and 12 bit gray-level resolution. For mass detection, the images were then subsampled (pixel digital value average) by a factor of 4 in both directions to generate images of approximately 600×450 pixels. All true-positive masses depicted in these images were pathologically verified, and the locations of the masses were marked on the images by radiologists.

Each image was processed by a multilayer topographic-based CAD scheme previously developed in our laboratory.[33] Each mammogram was processed as follows: Using dual-kernel filtering, subtraction, and simple thresholding methods, the scheme identifies a large number of suspicious mass regions. A set of image features is then extracted from the mammogram, and a classifier (i.e., artificial neural network) is applied to assign the region as a positive or negative one. In brief, this scheme has three distinct stages for the identification of masses. The first stage of dual kernel filtering, subtraction, and labeling resulted in the selection of a large number of suspicious regions (24 067 and 19 154 regions when applied to the two image databases, respectively, or approximately 24 regions per image). Based on local contrast measurements, the second stage used an adaptive region growth algorithm to define three topographic layers for each suspicious region. For each growth layer, a set of simple intralayer boundary conditions on region growth ratio and shape factor was applied to eliminate a large number of initial suspicious regions. After the second stage, the number of suspicious regions (including both positive and negative regions) decreased to 4382 and 3623 (or approximately 4.4 regions per image) in the training and testing databases. For each suspicious region, a set of image features was automatically computed by the scheme. Using these features, the third stage of the CAD scheme used a three-layer feed-forward ANN to classify these regions as positive or negative for mass.[24]

The second stage of the scheme identified 592 and 358 suspicious regions that depicted verified masses in the training and testing databases, respectively. With the exception of these regions that matched verified masses, all other regions that were identified as suspicious by the scheme at this stage were determined to be negative. A total of 3790 and 3265 negative regions were identified as suspicious (or false-positive) in the training and testing databases, respectively. For each region, 36 image features inside the suspicious region (including its three topographic growth layers[33]) and its surrounding background were automatically computed by the CAD scheme. These features include mainly geometrically related features, such as region size, circularity, or normalized standard deviation of radial length and intensity-related features (or distribution of pixel values), such as contrast, standard deviation, and skewness of pixel values' distribution and conspicuity. The definitions and the methods of computation for these features have been reported in several previous studies.[22,24] To reduce the potential redundancy and improve the robustness of the scheme, we used a genetic algorithm (GA) to select an optimal subset of input features to be used in the ANN.

B. Database segmentation

The basic concept of adaptive filtering is to divide suspicious regions (or images) into several groups based on a computable index and then to optimize different ANNs for the regions (or images) in each group. Although several complicated indices have been used for segmentation with some success,[26,27] we searched here for new indices. The selection criteria were: (1) the index was easily computable; (2) the index had been used as a feature in other CAD schemes; and (3) the relationship between the index and the segmentation results is "interpretable" and has been demonstrated in previous studies. Three indices were selected empirically for this study. The first is the size of the suspected region mul-

TABLE I. The number of false-positive regions in the training data set segmented by each of the indices into the "easy," "moderately difficult," and "difficult" groups, respectively.

Segmentation index	"Easy"	"Moderately difficult"	"Difficult"
Size×contrast	454	1002	2334
Conspicuity	227	741	2822
Circularity	366	849	2575

tiplied by its digital value contrast. This index could be interpreted to represent the "volume" of a suspicious mass. Studies have indicated that suspicious mass regions with large size and high contrast are easier to identify using CAD schemes than small regions with lower contrast.[25,34] The second index is region conspicuity. This index has been extensively investigated for the detection of lung nodules on chest images.[35] Radiologists typically achieved better diagnostic performance in detecting lung nodules with higher conspicuity than those with lower conspicuity.[36] A similar relationship between CAD performance and conspicuity of mass regions has also been demonstrated.[37] The third index is the region circularity, an important feature in classifying suspicious mass regions in a variety of CAD schemes.[24,38]

Using each of these indices, we divided suspicious regions into three groups, which were defined as "easy," "moderately difficult," and "difficult" regions. In order to have the same number of true-positive training samples in each of the three groups, two segmentation thresholds were determined based on the distribution of the feature values for the true-positive regions. As a result, the "easy" group included 198 true-positive regions, and the other two groups had 197 true-positive regions. The number of false-positive regions that resulted from such segmentation is listed in Table I. The same thresholds were applied later to the testing database.

C. GA optimization

In each group, a different classifier was used on the cases with similar characteristics. To search for an optimal set of features to apply to each group, a genetic algorithm (GA) was used. The binary coding method was applied to create a chromosome used in the GA. Each extracted feature corresponded to a gene. To decide the number of hidden neurons in the second (hidden) layer of the ANN, we added four genes in the chromosome. The chromosome had a fixed length of 40, where the first 36 genes represent extracted image features, and the last 4 genes indicate the number of hidden neurons. The same GA software and initial setup parameters have been reported previously.[22] In brief, the initial population size of chromosomes was set at 100. The crossover rate, the mutation rate, and the generation gap were set at 0.6, 0.001, and 1.0, respectively.

A training sample of equal number of true-positive and false-positive regions was then used to train the weights connecting the neurons in the ANN. To minimize the over-fitting and keep the robustness of ANN performance when applied to new cases, a limited number of training iterations as well as a large ratio between the momentum and learning rate was adopted.[24,39] The number of training iterations of the ANN was fixed at 1000, while the momentum and learning rate in the ANN training were set up as 0.8 and 0.01, respectively. ROC curves generated from the training samples (A_z values computed by the program ROCFIT)[40] were used as a fitness function (or criterion) in the GA optimization. The chromosomes that produced higher A_z values had higher probabilities of being selected in generating new chromosomes for the next generation using the methods of crossover and mutation. The GA was terminated when it converged to the highest A_z value or reached a predetermined number of generations (i.e., 100). The resulting set of features was assumed to be "optimal" and was implemented in the CAD scheme.

D. Adaptive and nonadaptive optimization

In this study we compared the performance changes of detection accuracy between the ANNs when optimized adaptively versus nonadaptively. In the adaptive optimization method, the training database was first segmented into three subsets with a "similar" characteristic. ANNs with different topologies and input features were then optimized separately using the GA method for each subset. To train an ANN, all true-positive regions in the subset were used, and the same number of false-positive regions was also randomly selected from the larger dataset of false-positive regions in that group. Using the GA method an ANN was optimized specifically for this subset. Since three segmentation indices (size×contrast, conspicuity, and circularity) were used in this experiment, a total of nine subsets, hence ANNs were established (three subsets for each segmentation index and three indices of segmentation).

In the nonadaptive optimization, the cases were not segmented into subsets. Because the number of training samples could affect performance,[24] we used the GA method to optimize the ANN once with 198 randomly selected true-positive and 198 false-positive regions (ANN-1), then we repeated the procedure including all 592 true-positive regions in the training database and a randomly selected set of 592 false-positive regions (ANN-2).

After optimization, an independent database, which includes 358 masses and 3265 regions that had been identified as suspicious, but were actually negative, was used to evaluate and compare the performance of the adaptive and nonadaptive ANNs. To test the adaptive scheme, the program first segmented the database into subsets using the same indices developed for the training phase. The ANN results for all regions in the testing database were used to compute the area under ROC curves (A_z values) using the ROCFIT program.

E. Performance gain by averaging scores

Averaging ratings cases from different independent readings could improve the diagnostic accuracy.[41] Accuracy gains are strongly dependent on the number of observations (or schemes) and the correlation between observations. For

TABLE II. Correlation coefficients between cases assigned to different groups using the segmentation rules based on the three features (size ×contrast, conspicuity, and circularity).

Indices compared	TP regions in training database	FP regions in training database	TP regions in testing database	FP regions in testing database
ANN-1 to ANN-2	0.148	0.174	0.152	0.209
ANN-1 to ANN-3	0.022	−0.069	0.008	−0.004
ANN-2 to ANN-3	0.219	0.018	0.298	0.005

TABLE III. The number of true- and false-positive regions assigned to the different groups using the three segmentation indices when applied to the testing database.

Segmentation index	Group 1 true/false positives	Group 2 true/false positives	Group 3 true/false positives
Size×contrast	120/514	123/893	115/1890
Conspicuity	113/182	116/612	129/2503
Circularity	106/290	107/791	145/2216

example, by averaging the results from three observations, accuracy gains could range from 0 and 73.2% when the correlations range from 1 to 0.[41]

Similar to the multireader problem, we segmented the data set three times using each of the three segmentation features (size×contrast, conspicuity, and circularity). Each segmentation resulted in three subsets of cases. Note that a case segmented into group one ("easy") based on one feature (e.g., circularity) may be classified into group three ("difficult") based on another feature (e.g., conspicuity). Each suspicious region was assigned to a specific category using each segmentation index, and the "optimal" ANN for that subset was applied by assigning a likelihood score. Hence, each region was assigned three different scores related to its likelihood for depicting a true mass. These scores were averaged and a "combined" ROC curve was generated. Results were compared to those obtained using individual scores. In addition, we compared experimentally measured and expected gains due to averaging based on measured correlations

$$\left(\rho_{X,Y}=\frac{\mathrm{COV}(X,Y)}{\sigma_X\sigma_Y}\right)$$

where $\mathrm{COV}(X,Y)$ is the covariance of two vectors X and Y, and σ_X and σ_Y are the standard deviations of the vectors, respectively.[42] The theoretical expected gains were computed for the averaging of multiple observations.[41]

III. RESULTS

Table I summarizes the number of false-positive regions assigned to each group when different features were used for segmentation in the training data set. Noted is the large number of regions assigned to the last "difficult" group. In general, this indicates that many of the false-positive regions were not "easy" to rule out as a true mass. The correlation coefficients between the classification assignment of regions based on the segmentation performed using the three features are summarized in Table II. The low correlations indicate that a large number of regions in each database were segmented into different groups when different features were used for segmentation. Only 12.5% of the true-positive regions and 25.2% of the false-positive regions in the training database were consistently assigned to the same group (e.g., easy). As a result, for the same training database, three sets of adaptive ANNs were actually trained with different cases for each group. When ANN scores from randomly selected groups with the same number of cases are compared, the correlation coefficients range from 0.712 to 0.963. These results clearly demonstrate that additional information could be obtained from the adaptive approach.

Table III provides the distribution of regions segmented into the different groups using the three segmentation indices in the testing database. While the percentage of large size×contrast regions ("easy" regions) is somewhat higher than that assigned to this group in the training database, the general distributions are quite similar. The optimization process resulted in ANNs that included different input features and varying numbers of hidden neurons. The number of input features ranged from 9 to 15 and the number of hidden neurons ranged from 3 to 7. Table IV provides the results (A_z) for the different schemes when applied to the testing database and a comparison (P values) to the nonadaptive scheme using 198 positive and 198 negative regions for training (ANN-1). The approach in ANN-2 is similar to ANN-1, only 592 positive and 592 negative regions were used for training purposes. Both ANN-1 and ANN-2 are nonadaptive schemes, and the significant improvement ($P=0.03$) in ANN-2 is largely the result of more complete feature domain coverage. Adaptive schemes 1–3 are the results after optimization by segmentation based on individual indices. For example, scheme 1 was trained using the subsets of size×contrast as a segmentation index. As can be seen, the results are somewhat better (albeit, not significantly) than the nonadaptive scheme using 198 positive and 198 negative regions (ANN-1), but these are not improved compared with ANN-2. On the other hand, by averaging detection scores of the different adaptive schemes (either two or all three), sig-

TABLE IV. Areas under ROC curves (A_z values) for different schemes and their comparisons (two-tailed p values) with the nonadaptive scheme using 198 positive and 198 negative regions (ANN-1).

Scheme	A_z[a]	P
Nonadaptive ANN-1	0.82	
Nonadaptive ANN-2	0.85	0.03
Adaptive-1	0.84	0.18
Adaptive-2	0.83	0.63
Adaptive-3	0.84	0.21
Average (1+2)	0.91	<0.01
Average (1+3)	0.92	<0.01
Average (2+3)	0.91	<0.01
Average (1+2+3)	0.95	<0.01

[a]Standard deviation for all A_z values is 0.01.

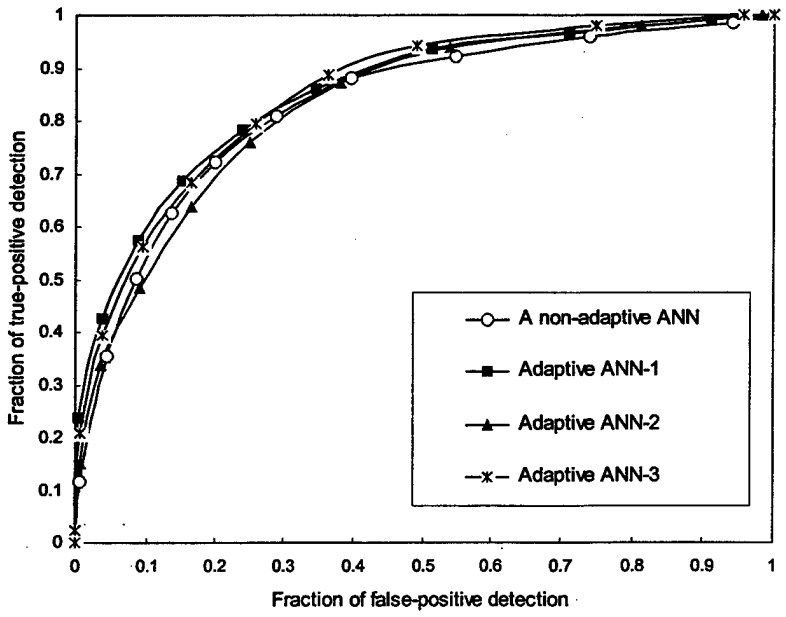

FIG. 1. ROC curves from nonadaptive ANN-1 and three sets of noncombined adaptive ANNs. The A_z values for these curves are 0.82, 0.84, 0.83, and 0.84, respectively.

nificant gains in detection accuracy ($p<0.01$) are achieved. Averaging results from two or three adaptive schemes resulted in a much larger performance gain ($P<0.01$) in the testing database as compared with ANN-2. Figures 1 and 2 demonstrate the ROC curves for several different classification schemes.

To verify the theoretical feasibility of obtaining the performance gains observed in this study, we used the correlations for the test results from the different adaptive schemes (Table V) in the estimation method proposed by Swensson *et al.*[41] to compute expected improvements by averaging these schemes. Table VI summarizes the predicted Z values and percentage gain in accuracy by averaging scores of two or three adaptive schemes. Predicted A_z values using a general binormal model are also provided. These are consistent with the experimental results we computed directly using ROCFIT.

IV. DISCUSSION

Averaging diagnostic ratings from different readers[41] or scores from different machine learning classifiers[17,21] might significantly improve detection accuracy, if the ratings or scores from different observations have low correlations. ANN is one of the most commonly used machine learning classifiers in CAD developments, due to its ability to learn complex patterns directly from training samples with minimal requirement on prior knowledge of the input features or internal system operation.[43] In this study, we explored a simple and novel method to segment and optimally train sets of adaptive ANNs. Since these produced extremely low correlated classification results using a large and independent testing database, significant gains were realized by averaging the scores from the different ANNs.

Given the large number of independent variables that are

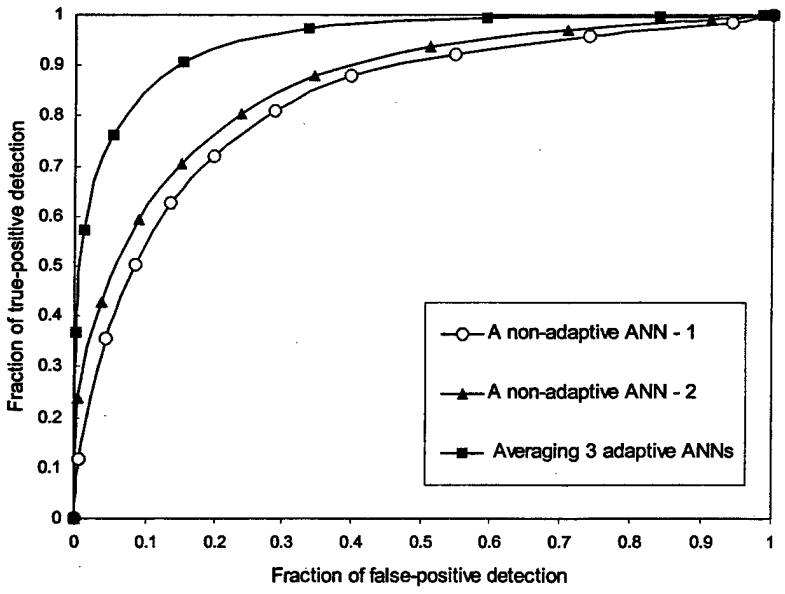

FIG. 2. ROC curves of classification results from nonadaptive schemes (ANN-1 and ANN-2) as well as after averaging scores of three sets of adaptive ANNs. The A_z values are 0.82 ± 0.01, 0.85 ± 0.01, and 0.95 ± 0.01, respectively.

TABLE V. Correlation coefficients between testing results using adaptive ANN scores from different schemes

Between adaptive schemes	TP regions [$\rho(a)$]	FP regions [$\rho(n)$]	Between A_z
ANN-1 to ANN-2	0.018	−0.004	0.013
ANN-1 to ANN-3	−0.011	0.003	−0.007
ANN-2 to ANN-3	0.116	0.011	0.086

needed to characterize masses and normal tissue structure on digitized mammograms and the fact that many of the features are continuous and span a wide range of values, a large and carefully selected training data set is required to ensure adequate domain coverage that could result in robust performance.[24] Finding an optimal feature set from a limited image database is an important factor in determining the performance and robustness of CAD schemes.[44,45] Had it been possible to extract an "ideal" (or fully optimized) set of features that adequately covers the variables' domain from a limited data set, it may not be necessary to perform the adaptive filtering and score averaging procedures described here. Using different training samples to optimize ANNs could result in different topologies (similar to using different input features or having different numbers of hidden neurons). However, our experiments showed that generally the correlations of the detection results when applying these ANNs to an independent testing database were quite high ($\rho \geq 0.7$).

In order to take advantage of possible improvement in performance due to score averaging, one should train different ANNs using the samples with different characteristics. The adaptive concept reported in previous CAD studies[26,27] was used here to group images with similar characteristics. The three segmentation indices reported in this study resulted in 87% of true-positive and 74% of false-positive regions being classified in different groups. Hence, the ANNs for the "same" group (e.g.,"easy" group) were trained using different images in each of the subsets segmented based on values from one of the three features. As a result, the classification scores generated by these three ANNs had low correlations. Similar to averaging ratings from independent observers,[28,29,41] averaging the scores from these "independent" ANNs yielded significant performance gains.

Although quite encouraging, the results presented here are preliminary and have to be validated in larger independent databases. We explored here only three simple and commonly used features for segmentation purposes. Other features, including those extracted locally (from a suspicious region) and globally (from a full image), should be explored as well. However, based on the results of this preliminary experiment, we believe that the approach taken may have significant advantages over a multifeature, single ANN approach to the problem.

TABLE VI. The predicted performance gain of averaging scores from the three adaptive schemes using the methodology proposed by Swensson *et al.* (Ref. 41).

Averaging adaptive schemes	Predicted Z (average)	Percentage gain in Z value	Predicted A_z	Measured A_z
1+2	1.374	48.2	0.92	0.91±0.01
1+3	1.420	53.1	0.92	0.92±0.01
2+3	1.338	44.3	0.91	0.91±0.01
1+2+3	1.644	77.3	0.95	0.95±0.01

ACKNOWLEDGMENTS

This work is supported in part by the U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick, MD 21702-5014 under Contract Nos 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 No. CA77850 from the National Cancer Institute, National Institutes of Health. The authors wish to thank William Reinus, M.D., and the research group at Washington University Medical School, St. Louis, MO, for providing some of the images used in this study.

[a]Electronic mail: zhengb@msx.upmc.edu

[1] W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology **191**, 331–337 (1994).

[2] H. P. Chan, S. C. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," Med. Phys. **22**, 1555–1567 (1995).

[3] L. Li, W. Qian, and L. P. Clarke, "Computer-assisted diagnosis method for mass detection with multiorientation and multiresolution wavelet transforms," Acad. Radiol. **4**, 724–731 (1997).

[4] W. E. Polakowski, D. A. Cournoyer, and S. K. Rogers, "Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency," IEEE Trans. Med. Imaging **16**, 811–819 (1997).

[5] A. J. Mendez, P. G. Tahocas, and M. J. Loda, "Computer-aided diagnosis: Automatic detection of malignant masses in digitized mammograms," Med. Phys. **25**, 957–964 (1998).

[6] W. Zhang, H. Yoshida, R. M. Nishikawa, and K. Doi, "Optimally weighted wavelet transform based on supervised training for the detection of microcalcifications in digital mammograms," Med. Phys. **25**, 949–956 (1998).

[7] H. D. Cheng, Y. M. Lui, and R. I. Freimanis, "A novel approach to microcalcification detection using fuzzy logic technique," IEEE Trans. Med. Imaging **17**, 442–450 (1998).

[8] S. Yu, L. Guan, and S. Brown, "Automatic detection of clustered microcalcifications in digitized mammogram films," J. Electron. Imaging **8**, 76–82 (1999).

[9] M. A. Gavrielides, J. Y. Lo, R. Vargas-Voracek, and C. E. Floyd, "Segmentation of suspicious clustered microcalcifications in mammograms," Med. Phys. **27**, 13–22 (2000).

[10] B. Zheng, J. H. Sumkin, W. F. Good, G. S. Maitz, Y. H. Chang, and D. Gur, "Applying computer-assisted detection schemes to digitized mammograms after JPEG data compression: An assessment," Acad. Radiol. **7**, 595–602 (2000).

[11] C. J. Vyborny and M. L. Giger, "Computer vision and artificial intelligence in mammography," AJR, Am. J. Roentgenol. **162**, 699–708 (1994).

[12] K. R. Hoffman, "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. **26**, 1–2 (1999).

[13] L. J. Burhenne, S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. F. O'Shaughnessy, E. A. Sickles, L. Tabar, C. J. Vyborny, and R. A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," Radiology **215**, 554–562 (2000).

[14]G. M. Brake, N. Karssemeijer, and J. H. Hendriks, "Automated detection of breast carcinomas not detected in a screening program," Radiology **207**, 465–471 (1998).
[15]J. E. Gray, "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. **26**, 3–4 (1999).
[16]D. Wei, H. P. Chan, N. Pertrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "False-positive reduction technique for detection of masses on digital mammograms: Global and local multiresolution texture analysis," Med. Phys. **24**, 903–914 (1997).
[17]R. H. Nagel, R. M. Nishikawa, J. Papaioannou, and K. Doi, "Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms," Med. Phys. **25**, 1502–1506 (1998).
[18]B. Sahiner, H. P. Chan, D. Wei, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue," Med. Phys. **23**, 1671–1684 (1996).
[19]M. A. Anastasio, H. Yoshida, R. Nagel, R. M. Nishikawa, and K. Doi, "A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms," Med. Phys. **25**, 1613–1620 (1998).
[20]W. Zhang, K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, "An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms," Med. Phys. **23**, 595–601 (1996).
[21]R. Rymon, B. Zheng, Y. H. Chang, and, D. Gur, "Incorporation of a set enumeration tree-based classifier into a hybrid computer-assisted diagnosis scheme for mass detection," Acad. Radiol. **5**, 181–187 (1998).
[22]B. Zheng, Y. H. Chang, X. H. Wang, W. F. Good, and D. Gur, "Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm," Acad. Radiol. **6**, 327–332 (1999).
[23]Y. H. Chang, L. A. Hardesty, C. M. Hakim, T. S. Chang, B. Zheng, W. F. Good, and D. Gur, "Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment," Med. Phys. **28**, 455–461 (2001).
[24]B. Zheng, Y. H. Chang, W. F. Good, and D. Gur, "Adequacy testing of training set sample size in the development of a computer-assisted diagnosis scheme," Acad. Radiol. **4**, 497–502 (1997).
[25]R. M. Nishikawa, M. L. Giger, K. Doi, C. E. Metz, F. F. Yin, C. J. Vyborny, and R. A. Schmidt, "Effect of case selection on the performance of computer-aided detection schemes," Med. Phys. **21**, 265–269 (1994).
[26]B. Zheng, Y. H. Chang, and D. Gur, "Adaptive computer-aided diagnosis scheme of digitized mammograms," Acad. Radiol. **3**, 806–814 (1996).
[27]W. Qian, L. Li, L. Clarke, R. A. Clark, and J. Thomas, "Digital mammography: Comparison of adaptive and nonadaptive CAD schemes for mass detection," Acad. Radiol. **6**, 471–480 (1999).
[28]C. E. Metz and J. H. Shen, "Gains in accuracy from replicated readings of diagnostic images: Prediction and assessment in terms of ROC analysis," Med. Decis. Making **12**, 60–75 (1992).
[29]R. G. Swensson and P. F. Judy, "Measuring performance efficiency and consistency in visual discriminations with noisy images," J. Exp. Psychol. **22**, 1393–1415 (1996).
[30]S. Nawano, K. Murakami, N. Moriyama, and H. Kobatake, "Computer-aided diagnosis in full digital mammography," Invest. Radiol. **34**, 310–316 (1999).
[31]T. Doi, A. Hasegawa, B. Hunt, J. Marshall, F. Rao, and J. Roehrig, "Clinical results with the R2 ImageCheck Mammographic CAD system," in *Computer-aided Diagnosis*, edited by K. Doi, H. MacMahon, M. L. Giger, and K. R. Hoffman (Elsevier, Amsterdam, 1999), pp. 201–207.
[32]B. Zheng, M. A. Ganott, C. A. Britton, C. M. Hakim, L. A. Hardesty, T. S. Chang, H. E. Rockette, and D. Gur, "Soft-display mammographic readings under different computer-assisted detection cueing environments: Preliminary findings," Radiology (in press).
[33]B. Zheng, Y. H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single-image segmentation and a multiplayer topographic feature analysis," Acad. Radiol. **2**, 959–966 (1995).
[34]R. M. Nishikawa and L. M. Yarusso, "Variations in measured performance of CAD schemes due to database composition and scoring protocol," Proc. SPIE **3338**, 840–844 (1998).
[35]H. L. Kundel and G. Revesz, "Lesion conspicuity, structure noise, and film reader error," AJR, Am. J. Roentgenol. **126**, 1233–1238 (1976).
[36]G. Revesz, H. L. Kundel, and L. C. Toto, "Densitometric measurements of lung nodules on chest radiographs," Invest. Radiol. **16**, 201–205 (1981).
[37]B. Zheng, Y. H. Chang, W. F. Good, and D. Gur, "Assessment of mass detection using tissue background information as input to a computer-assisted diagnosis scheme," Proc. SPIE **3338**, 1547–1555 (1998).
[38]M. Kupinski, M. L. Giger, P. Lu, and Z. M. Huo, "Computerized detection of mammographic lesions: Performance of artificial neural network with enhanced feature extraction," Proc. SPIE **2434**, 598–605 (1995).
[39]B. Zheng, W. F. Good, X. H. Wang, and Y. H. Chang, "Comparison of artificial neural network and Bayesian belief network in a computer-assisted diagnosis scheme for mammography," Proceedings of the International Joint Conference on Neural Network, Washington, DC, 10–16 July, 1999.
[40]C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," Stat. Med. **17**, 1033–1053 (1998).
[41]R. G. Swensson, J. L. King, W. F. Good, and D. Gur, "Observer variation and the performance accuracy gained by averaging ratings of abnormality," Med. Phys. **27**, 1920–1933 (2000).
[42]A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering* (Addison–Wesley, Reading, MA, 1994), p. 233.
[43]J. Diederich, "Explanation and artificial neural networks," Int. J. Man-Mach. Stud. **37**, 335–341 (1992).
[44]M. A. Kupinski and M. L. Giger, "Feature selection with limited databases," Med. Phys. **26**, 2176–2182 (1999).
[45]H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," Med. Phys. **26**, 2654–2668 (1999).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

MATERIALS AND METHODS

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

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

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

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

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

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

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

TABLE 1
Number of Mammographic Cases in Different Categories

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

Note.—B = benign, M = malignant.

TABLE 2
CAD Cuing Conditions of the Five Display Modes

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

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

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

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

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

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

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

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

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

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

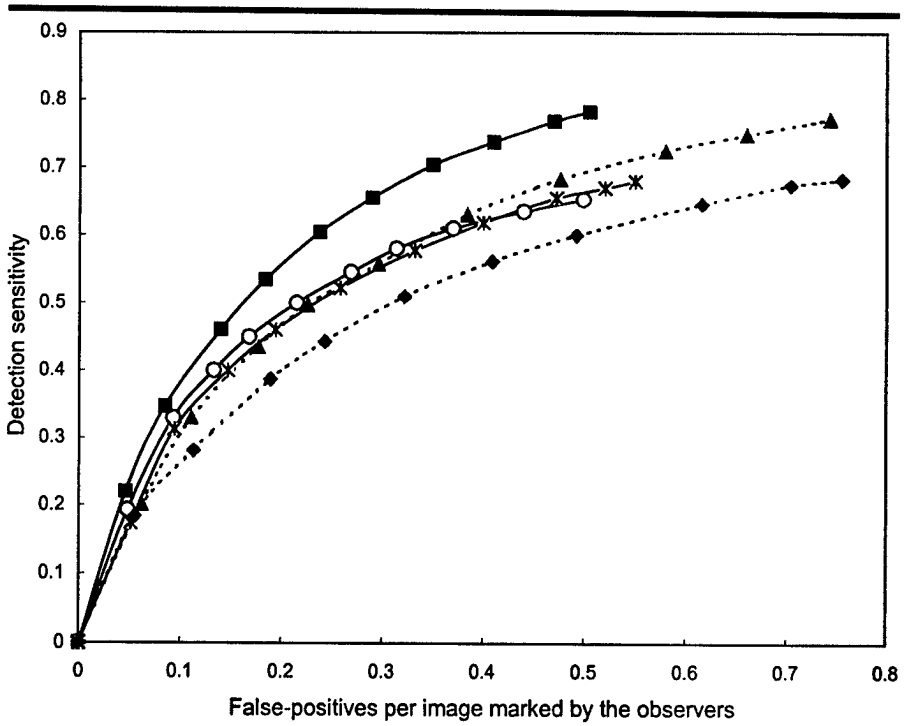

Figure 1. Free-response receiver operating characteristic curves for the average detection of mammographic abnormalities (including both masses and microcalcification clusters) by seven participating radiologists using five display modes. ○ = mode 1, ■ = mode 2, ▲ = mode 3, * = mode 4, and ◆ = mode 5.

RESULTS

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

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

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

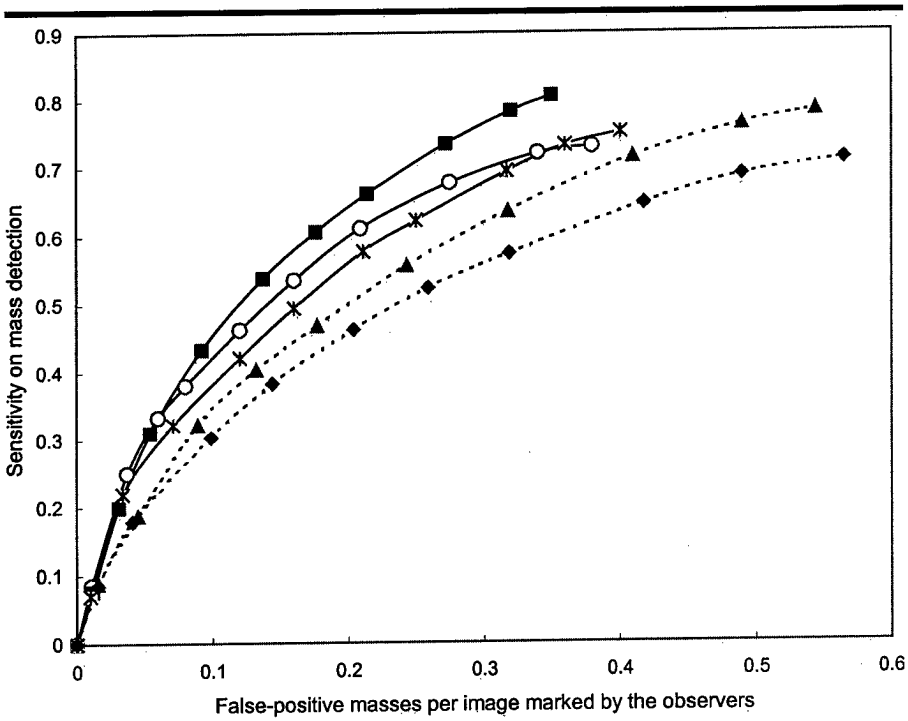

Figure 2. Free-response receiver operating characteristic curves for the average mass detection by seven radiologists using five display modes. ○ = mode 1, ■ = mode 2, ▲ = mode 3, ∗ = mode 4, and ◆ = mode 5.

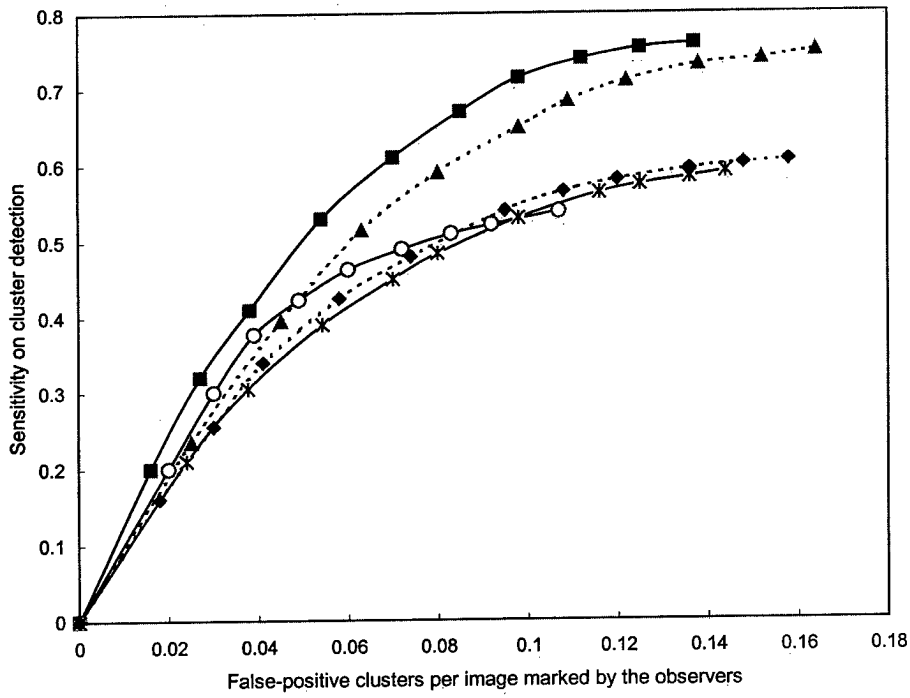

Figure 3. Free-response receiver operating characteristic curves for the average microcalcification cluster detection by seven radiologists using five display modes. ○ = mode 1, ■ = mode 2, ▲ = mode 3, ∗ = mode 4, and ◆ = mode 5.

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

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

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

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

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

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

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

TABLE 4
Number of Missed Abnormalities in Noncued Regions

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

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

DISCUSSION

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

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

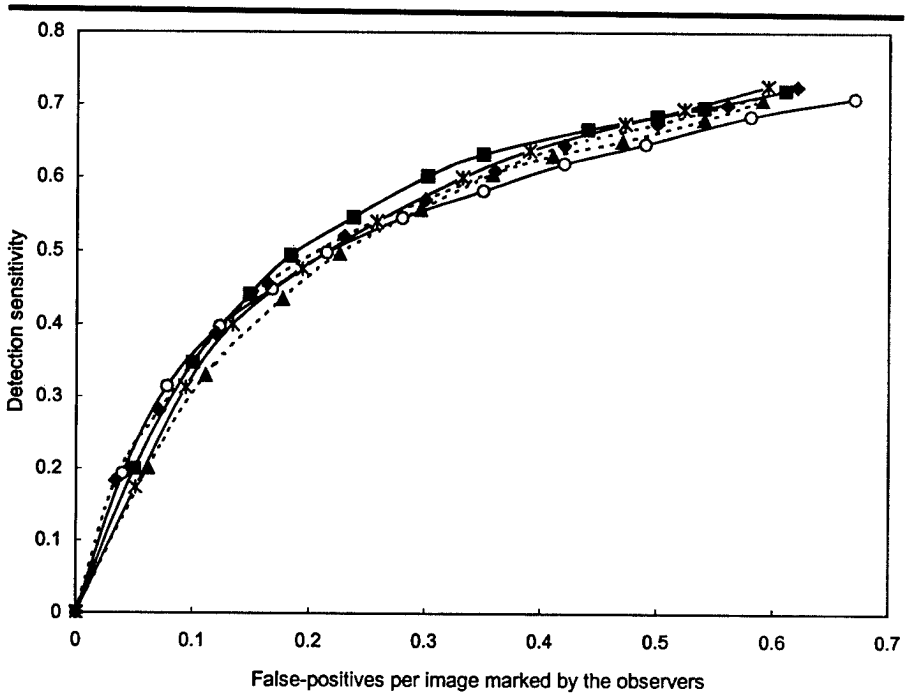

Figure 4. Free-response receiver operating characteristic curves for the average detection of abnormalities by seven radiologists as a function of the order of appearance: ○ = first time, ■ = second time, ▲ = third time, * = fourth time, and ◆ = fifth time, regardless of the reading mode.

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

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

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

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

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

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

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

References

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

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

dhl
11-006
8-9-02

AUTHOR

[artno]11006

[vol]9

[iss]11

[mo]November

[yr]2002

[lrh]Zheng et al[/lrh]

[rrh]CAD ON CURRENT AND PRIOR IMAGES[/rrh]

[sh]Original Investigations[/sh]

[t]Computer-aided Detection in Mammography: An Assessment of Performance on Current and Prior Images[1][fn1][/t]

[aut]Bin Zheng, PhD, Ratan Shah, MD, Luisa Wallace, MD, Christiane Hakim, MD, Marie A. Ganott, MD, David Gur, ScD[/aut]

Rationale and Objectives. The authors assessed and compared the performance of a computer-aided detection (CAD) scheme for the detection of masses and microcalcification clusters on a set of images collected from two consecutive ("current" and "prior") mammographic examinations.

Materials and Methods. A previously developed CAD scheme was used to assess two consecutive screening mammograms from 200 cases in which the current mammogram showed a mass or cluster of microcalcifications that resulted in breast biopsy.<a> The latest prior examinations had been initially interpreted as negative or

---

[a]Au: Verify or correct rewording. The quotation marks around "current" and "prior" have been dropped after the first mention. OK

definitely benign findings (Breast Imaging Reporting and Data System rating, 1 or 2). The study involved images of 400 examinations acquired in 200 patients. Radiologists identified 172 masses and 128 clusters of microcalcifications on the current images. The performance of the CAD scheme was analyzed and compared for the current and latest prior images.

Results. There were significant differences ($P < .01$) between current and prior images in many feature values. The performance of the CAD scheme was significantly lower for prior than for current images ($P < .01$). At 0.5 and 0.2 false-positive mass and cluster cues per image, the scheme detected 78 malignant masses (78%) and 63 malignant clusters (80%) on current images. Only 42% of malignant cases were detected on prior images, including 40 masses (40%) and 36 microcalcification clusters (46%).

Conclusion. CAD schemes can ~~demonstrate a significant~~ detect a substantial <b> fraction of masses and microcalcification clusters depicted on prior images. To improve performance with prior images, the scheme may have to be adaptively reoptimized with increasingly more subtle abnormalities.

Key Words. Breast, calcification; breast neoplasms, diagnosis; breast radiography; computers, diagnostic aid.

[fnote][fn][1]From the Department of Radiology, Ste 4200,

---

[b]Au: Please reword "significant," unless you mean (statistically) significantly more. Change "depicted" to "depicted but not identified"? see correction. "demonstrate" is not correct.

University of Pittsburgh and Magee-Womens Hospital, 300 Halket St, Pittsburgh, PA 15213. Received July 16, 2002; accepted July 17. Supported in part by grants CA77850, CA85241, and CA80836 from the National Cancer Institute of the National Institutes of Health and by the U.S. Army Medical Research Acquisition Center under Contract DAMD17-00-1-0410. **Address correspondence to** B.Z.

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

Breast cancer is a common cancer in women over the age of 40 years (1). Early detection is believed to be important for improved prognosis and therapy and for reducing associated mortality and morbidity (2). Mammography is a well-established and accepted method for screening the general population. Current guidelines in the United States recommend periodic mammographic screening for women aged 40 years or older (3). Because of the large volumes, low expected detection rate of abnormalities in screening examinations, and the complexity of tissue patterns depicted on a large fraction of images, it is both difficult and time consuming to interpret mammographic cases (4). Independent double reading is a well-documented method to improve early detection of breast cancer (5,6), but this approach is often not practical due to personnel and logistic constraints (7).

After extensive investigations and development efforts for more than a decade, computer-aided detection (CAD) systems have been accepted as clinical tools that provide radiologists with a useful "second opinion." Three CAD systems, ImageChecker (R2 Technology, Los Altos, Calif), Second Look (CADx Medical Systems, Quebec, Canada), and MammoReader (Intelligent Systems Software, Clearwater, Fla) have been approved to date by the U.S. Food and Drug Administration for this purpose. Their performance has been evaluated (8–10). While in general the systems have been shown to increase sensitivity, these results are not universal. One study reported that, with the help of a commercially available system, two radiologists detected 19.5% more cancers with only a slight increase (from 6.5% to 7.7%) in recall rate (11). Another study reported that use of a comparable system did not affect the performance of three radiologists retrospectively interpreting a set of mammograms depicting 59 breast cancers in 280 patients (no increase in sensitivity or decrease in specificity) (12). Our own preliminary study, in which seven radiologists interpreted 120 mammographic cases under five different CAD cueing conditions, suggested that highly performing CAD schemes can significantly improve the diagnostic performance of radiologists, while poorly performing schemes can adversely affect performance (13).

One objective of using CAD is the potential to detect breast cancers at an earlier stage. It is well known that a large number

of breast abnormalities (ie, masses and microcalcification clusters) are visible in retrospect on prior mammograms but are not interpreted at the time as highly suspicious. In one study, 427 breast cancer cases were reviewed, and the abnormality in question was visible on the latest prior mammograms in 286 (67%) (9). When 115 of the "more obvious" cases (27% of the original 427 cases) were processed by a CAD system, 89 cancers (or 77%) were identified as suspicious on the prior mammograms, with an average of one false-positive cue per image (14). Commercial systems generally provide only a binary outcome for each suspicious region (cued or not cued) based on a predetermined (and undisclosed) threshold. Therefore, the difference in performance between different groups of images (in this case "current" and "prior") can be measured only at one operating point. Hence, complete characterization (eg, a free-response receiver operating characteristic [FROC]-type curve) of the performance cannot be estimated (8,14).

In the study reported here, we applied a CAD scheme previously developed in our laboratory to a set of 200 selected cases with mammograms from two consecutive examinations. At the latest examination (current images), at least one suspicious mass or microcalcification cluster was identified by the interpreting radiologist, resulting in breast biopsy. For the prior examinations, all images were interpreted as "negative" or

"benign finding."

MATERIALS AND METHODS

The mammographic cases used in this study were selected from biopsy records of two medical facilities in Pittsburgh, Pa. In one facility we collected all available biopsy cases performed in 1997, and in another we ascertained a fraction of the biopsy cases performed in 2000. First, we excluded cases for which all the original mammograms from the latest prior examination were not available. Second, we excluded cases in which the recommendations for biopsy had not been based on either the finding of mass or microcalcification cluster. Third, we selected only cases whose findings had been interpreted as either negative or benign (Breast Imaging Reporting and Data System<c> rating on the latest prior examination, 1 or 2).

From the remaining pool, 200 cases were selected sequentially for the study. Each case included images acquired from two consecutive examinations. In this set of 200 cases, the interval between the current examination (when the patient was sent to biopsy) and the latest prior examination varied from 10 to 22 months. Radiologists identified 172 masses and 128 microcalcification clusters in this data set. Of the 172 identified masses, 164 were visible (in retrospect) on both views

---

[c]Au: Addition OK, as in the abstract? OK

(craniocaudal [CC] and mediolateral oblique [MLO]),[d] and eight were visible only on one view. One hundred twenty of 128 microcalcification clusters were visible on two views, and eight on only one. Hence, there were a total of 336 mass regions and 248 cluster regions depicted on these mammograms. One hundred masses and 79 clusters were associated with malignancies. Two masses and four clusters were visible on only one view. Therefore, 198 mass regions and 154 cluster regions depicted on the current images were associated with malignancy. Table 1 summarizes the distributions of abnormalities by type and abnormality in the database. A fraction of the masses and clusters were visible on the prior images. Therefore, the corresponding locations of all mass and cluster regions on prior images were determined visually during a side-by-side inspection and after differences in breast positioning and compression were accounted for subjectively.

All mammograms were digitized in our laboratory with a laser film digitizer (Lumisys; city, state[e]) with a pixel size of 50 × 50 μm and 12 bits of gray levels. Each image was then subsampled by a factor of two in both dimensions with a pixel averaging method to reduce the spatial resolution to 100 × 100 μm. Our previously described CAD scheme (15) was applied to the

EASTMAN KODAK; ROCHESTER, NY

---

[d]Au: Verify expansions of CC and MLO. Yes

[e]Au: Please provide the manufacturer's location.
Lumisys digitizers are now manufactured by The Eastman Kodak Company, Rochester, NY

images to detect suspicious regions for microcalcification clusters. Images were then subsampled again by a factor of four in both dimensions to reduce the effective pixel size to 400 × 400 μm, and a "mass" detection scheme (16) was applied.

The CAD scheme developed in our laboratory (15-17) was applied without modifications ("as is") to all images in the database. After image segmentation and topographic multilayer region growth (15,16), the scheme extracts a set of image features for each identified suspicious region and its surrounding tissue background. Two artificial neural networks (ANNs), one for mass detection and one for microcalcification cluster detection, were used to classify each suspicious region by assigning it a likelihood score for the abnormality in question (for the likelihood of being positive) (17). With these detection scores used as the input values of an ROC curve-fitting routine (18), performance curves were generated. After normalization for the maximum false-positive rates, the performance results were transformed into FROC curves. FROC curves were compared for the corresponding current and prior image data sets.

False-positive cueing rates are extremely important in the screening environment (12,13). Therefore, in our analysis, we used as operating points false-positive rates of 0.5 per image for masses and 0.2 per image for microcalcification clusters,

similar to the reported performance levels of commercially available CAD systems (10,11) and our own experimental results (13). At these false-positive rates, we compared the detection sensitivities for masses and clusters between the current and prior images. For malignant mass and microcalcification cluster regions that were initially identified as suspicious by the CAD scheme on both current and prior images but were ultimately cued only on current images, we analyzed changes in the main features used in the ANN, to clarify why low output scores were generated for these regions on prior images (or why these were ultimately discarded by the scheme).[f]

Both "case-based" and "region-based" sensitivities were assessed in this study. Case-based sensitivity includes correct cues of an abnormality (eg, a mass or cluster) on one or both views (CC, MLO, or both); a "case" here means one abnormality and not necessarily one patient. Region-based sensitivity includes correct cues of an abnormality depicted independently on either view[g] (CC or MLO). The same abnormality depicted on both views (CC and MLO) is considered two independent true-positive findings. Region-based sensitivity was computed according to the number of correctly detected regions, rather than abnormalities.

---

[f]Au: Verify or correct rewording. OK

[g]Au: Verify "either view." Yes

RESULTS

Figures 1 and 2 demonstrate the case-based FROC curves for current and prior images for the detection of masses and microcalcification clusters, respectively. Figures 3 and 4 demonstrate the region-based FROC curves for mass and cluster detection. Figures 5 and 6 demonstrate FROC curves of case-based detection sensitivity versus false-positive rate for malignant mass and cluster detection, respectively, after the exclusion of biopsy-proved benign cases. The CAD scheme detected (though at a high false-positive rate) 94% of masses (162 of 172) and 95% of microcalcification clusters (122 of 128) in the current image database.

For the prior image database, the maximum detection sensitivities were 86% for masses (148 of 172) and 73% for clusters (93 of 128), as shown in Figures 1 and 2. After benign abnormalities were excluded, similar maximum sensitivities were obtained for mass and cluster detections: 95% for both masses (95 of 100) and clusters (75 of 79) on the current images and 76% (76 of 100) and 59% (47 of 79) for masses and clusters, respectively,[h] on prior images (Figs 5, 6). The scheme has comparable performance levels for detecting malignant or benign findings on current images. Its sensitivity for malignant lesions, however, is significantly lower than that for benign

---

[h]Au: Verify editing.

lesions on prior images ($P < .01$).

With specific thresholds set on the ANN-generated scores (0.55 for mass detection and 0.5 for cluster detection), the false-positive rates in our database were 0.5 per image for masses and 0.2 per image for microcalcification clusters. At these threshold levels, our CAD scheme detected 78% of malignant masses (78 cases or 109 regions) and 80% of malignant clusters (63 cases or 92 regions) on the current images. Suspicious regions that were cued in the corresponding areas of prior images were 53 "mass regions" (or 40 "masses") and 51 "cluster regions" (or 36 "clusters").<i> The case-based sensitivities for prior images were 40% (40 of 100) for malignant masses and 46% (36 of 79) for malignant clusters.

For mass detection, 24 malignant regions were cued on the current images but not on the prior images. In six features used in the ANN (17) for mass detection, the average feature values changed significantly ($P < .05$) between current and prior images. Table 2 summarizes the changes in these features. The estimated "size" and "contrast" of the cued regions were significantly smaller ($P < .05$) on prior images. In general, because of these changes, the mass regions depicted on prior images are more difficult to identify, not only for human observers but also for the CAD schemes optimized on a different set of cases (19,20).

---

[i]Au: Are the quotations marks necessary here? yes

For microcalcification detection, 21 malignant cluster regions were cued on the current images but not on the prior images due to lower ANN-generated scores. Of 13 features used in the ANN for cluster detection (17), only two had a significant change ($P < .05$) in average values between current and prior images. As may be expected, one was the number of single "microcalcifications" detected in a cluster, which was 25% smaller on prior images (5.6 per cluster vs 8.2 on current images).<j> The second was the average digital value contrast of a single microcalcification, which was 24% less on prior images.

DISCUSSION

There is a growing interest in using CAD to help detect breast cancers at an earlier stage. Hence, there is a need to detect some abnormalities depicted on prior images (9,14,21). In previous studies, CAD schemes were applied mainly to cases interpreted as recommended for recall by a panel of radiologists during retrospective reviews. In this study, we applied a CAD scheme to prior examinations of cases that ultimately underwent biopsy because of findings during a subsequent examination. Our experimental results showed that 76% of malignant masses and 59% of clusters associated with malignancies were detected as

---

[j]Au: Verify the rewording in this passage. Does "microcalcifications" need to be in quotation marks? No.

suspicious with the CAD scheme (Figs 5, 6). By applying thresholds on the ANN scores to generate false-positive rates of 0.5 per image for mass regions and 0.2 per image for cluster regions, the scheme ultimately detected 42% of cancers depicted on prior images. This is in the range of the fraction of cases reported to be visible at prior examinations in other studies (9).

The detection of abnormalities was found to be more sensitive to changes in feature values on the prior images. For example, reducing the false-positive rate for mass detection from 1.0 to 0.5 per image decreased sensitivity by 14% (from 0.88 to 0.76)[k] on the current images and 31% (from 0.58 to 0.40) on the prior images (Fig 5). Our experiment also suggested that the set of features that optimally represent malignant masses may be somewhat different on current and prior images (Table 2). This observation is in agreement with that in another study in which a stepwise linear discriminant analysis selected different sets of optimal features to represent masses depicted on current and prior images (22).

Unlike other studies using a commercial CAD product (8,14), for which only one operating point (detection sensitivity at a given false-positive rate) can be analyzed, this study generated

---

[k]Au: Verify expression of sensitivities as decimals, and note other editing in this paragraph. OK

complete FROC curves. Hence, one can compare the performance difference at any operating point and investigate the effect of feature changes on performance. This approach may represent an important first step toward reoptimizing CAD schemes that improve the detection of breast cancers at an earlier stage. Such early detection will become increasingly important, because the average stage at detection will gradually shift toward that seen on prior images as compliance improves and women undergo several periodic examinations.

Finally, full-field digital mammographic systems are rapidly becoming available (23,24). Although we did not include them in this study, we expect that the questions we considered are as relevant to full-field digital mammograms as to digitized film images.<1>

REFERENCES

1. Mettlin C. Global breast cancer mortality statistics. CA Cancer J Clin 1999; 49:135-137.

2. Rennie J, Rusting R. Making headway against cancer. Sci Am 1996; 3:56-59.

3. Feig SA, D'Orsi CJ, Hendrick RE. American College of Radiology guidelines for breast cancer screening. AJR Am J Roentgenol 1998; 171:29-33.

---

[1]Au: Verify rewording. OK

4. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology 1992; 184:613-617.

5. Thurfjell EL, Lernevall KA, Taube AS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241-244.

6. Hendee WR. Proposition: all mammograms should be double-read. Med Phys 1999; 26:115-117.

7. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. Acad Radiol 1996; 3:891-897.

8. Malich A, Azhari T, Bohm T, Fleck M, Kaiser WA. Reproducibility: an important factor determining the quality of computer aided detection (CAD) systems. Eur J Radiol 2000; 36:170-174.

9. 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.

10. Malich A, Marx C, Facius M, Boehm T, Fleck M, Kaiser WA. Tumour detection rate of a new commercially available computer-aided detection system. Eur Radiol 2001; 11:2454-2459.

11. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology 2001; 220:781-786.

12. Moberg K, Bjurstam N, Wilczek B, Rostgard L, Egge E,

Muren C. Computed assisted detection of interval breast cancers. Eur J Radiol 2001; 39:104-110.

13. Zheng B, Ganott MA, Britton CA, et al. Soft-copy mammographic reading with different computer-assisted detection cueing environments: preliminary findings. Radiology 2001; 221:633-640.

14. 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.

15. 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.

16. 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.

17. Zheng B, Sumkin JH, Good WF, Maitz GS, Chang YH, Gur D. Applying computer-assisted detection schemes to digitized mammograms after JPEG data compression. Acad Radiol 2000; 7:595-602.

18. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat Med 1998; 17:1033-1053.

19. Nishikawa RM, Giger ML, Doi K, Yin FF, Metz CE, Schmidt RA. Effect of case selection on the performance of computer-aided detection schemes. Med Phys 1994; 21:265-269.

20. 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. Med Phys 2001; 28:2302-2308.

21. Brake GM, Karssemeijer N, Hendriks JH. Automated detection of breast carcinomas not detected in a screening program. Radiology 1998; 207:465-471.

22. Hadjiiski L, Sahiner B, Chan HP, Petrick N, Helvie MA. Analysis of temporal changes of mammographic features: Computer-aided classification of malignant and benign breast masses. Med Phys 2001; 28:2309-2317.

23. Lewin JM, Hendric RE, D'Orsi CJ, et al. Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. Radiology 2001; 218:873-880.

24. Venta LA, Hendrick RE, Adler YT, et al. Rates and causes of disagreement in interpretation of full-field digital mammography and screen-film mammography in a diagnostic setting. AJR Am J Roentgenol 2001; 176:1241-1248.

Figure 1. Comparison of case-based CAD performance for detection

of masses on 200 current and prior mammographic cases. The test set included 172 masses.

Figure 2. Comparison of case-based CAD performance for detection of microcalcification clusters on 200 current and prior mammographic cases. The test set included 128 true-positive clusters.

Figure 3. Comparison of region-based CAD performance for detection of masses on current and prior images. The test set included 336 mass regions.

Figure 4. Comparison of region-based CAD performance for detection of microcalcification clusters on current and prior images. The test set included 248 cluster regions.

Figure 5. Comparison of case-based CAD performance for malignant mass detection. The test set included 100 malignant masses.

Figure 6. Comparison of case-based CAD performance for malignant microcalcification cluster detection. The test set included 79 malignant clusters.

Table 1
Distribution of Selected Masses and Microcalcification Clusters

Type of Abnormality	All Cases			Malignant Cases		
	Total	Visible on 2 Views	Visible on 1 View	Total	Visible on 2 Views	Visible on 1 View
Mass only	153	145	8	83	81	2
Cluster only	109	101	8	62	58	4
Mass and clusters combined	19	19	0	17	17	0

Table 2
Average Values of Six Features and Change in Values between Current and Prior Images for 24 Malignant Masses

Value	Region Size (mm^2)	Contrast (digital value)	Circularity	Standard Deviation of Radial Length	Pixel Ratio of Local Minimum Digital Value	Region Conspicuity
Average for current images	133.1 ± 100.2	42.1 ± 10.7	0.83 ± 0.07	0.21 ± 0.07	0.13 ± 0.05	4.7 ± 1.5
Average for prior images	66.3 ± 41.4	33.9 ± 12.3	0.76 ± 0.09	0.29 ± 0.08	0.21 ± 0.07	3.7 ± 0.7
Change (%)	−50.2	−19.5	−8.4	+38.1	+61.5	−21.3

Note.—These 24 masses were ultimately cued on the current images but not on the prior images ($P < .05$ for each of the six features). Mean values are given ± standard deviations.

Table 1
Distribution of Selected Masses and Microcalcification Clusters

Type of Abnormality	All Cases			Malignant Cases		
	Total	Visible on 2 Views	Visible on 1 View	Total	Visible on 2 Views	Visible on 1 View
Mass only	153	145	8	83	81	2
Cluster only	109	101	8	62	58	4
Mass and clusters combined	19	19	0	17	17	0

Table 2
Average Values of Six Features and Change in Values between Current and Prior Images for 24 Malignant Masses

Value	Region Size (mm^2)	Contrast (digital value)	Circularity	Standard Deviation of Radial Length	Pixel Ratio of Local Minimum Digital Value	Region Conspicuity
Average for current images	133.1 ± 100.2	42.1 ± 10.7	0.83 ± 0.07	0.21 ± 0.07	0.13 ± 0.05	4.7 ± 1.5
Average for prior images	66.3 ± 41.4	33.9 ± 12.3	0.76 ± 0.09	0.29 ± 0.08	0.21 ± 0.07	3.7 ± 0.7
Change (%)	−50.2	−19.5	−8.4	+38.1	+61.5	−21.3

Note.—These 24 masses were ultimately cued on the current images but not on the prior images ($P < .05$ for each of the six features). Mean values are given ± standard deviations.

Table 1
Distribution of Selected Masses and Microcalcification Clusters

Type of Abnormality	All Cases			Malignant Cases		
	Total	Visible on 2 Views	Visible on 1 View	Total	Visible on 2 Views	Visible on 1 View
Mass only	153	145	8	83	81	2
Cluster only	109	101	8	62	58	4
Mass and clusters combined	19	19	0	17	17	0

Table 2
Average Values of Six Features and Change in Values between Current and Prior Images for 24 Malignant Masses

Value	Region Size (mm^2)	Contrast (digital value)	Circularity	Standard Deviation of Radial Length	Pixel Ratio of Local Minimum Digital Value	Region Conspicuity
Average for current images	133.1 ± 100.2	42.1 ± 10.7	0.83 ± 0.07	0.21 ± 0.07	0.13 ± 0.05	4.7 ± 1.5
Average for prior images	66.3 ± 41.4	33.9 ± 12.3	0.76 ± 0.09	0.29 ± 0.08	0.21 ± 0.07	3.7 ± 0.7
Change (%)	−50.2	−19.5	−8.4	+38.1	+61.5	−21.3

Note.—These 24 masses were ultimately cued on the current images but not on the prior images ($P < .05$ for each of the six features). Mean values are given ± standard deviations.